Color Theme Evaluation through User Preference Modeling

BAILIN YANG, Zhejiang Gongshang University, Hangzhou, China
TIANXIANG WEI, Tianjin University, Tianjin, China
FREDERICK W. B. LI, University of Durham, Durham, United Kingdom
XIAOHUI LIANG, Beihang University, Beijing, China
ZHIGANG DENG, University of Houston, Houston, TX, USA
YILI FANG, Zhejiang Gongshang University, Hangzhou, China

Color composition (or color theme) is a key factor to determine how well a piece of art work or graphical design is perceived by humans. Despite a few color harmony models have been proposed, their results are often less satisfactory since they mostly neglect the variations of aesthetic cognition among individuals and treat the influence of all ratings equally as if they were all rated by the same anonymous user. To overcome this issue, in this article we propose a new color theme evaluation model by combining a back propagation neural network and a kernel probabilistic model to infer both the color theme rating and the user aesthetic preference. Our experiment results show that our model can predict more accurate and personalized color theme ratings than state of the art methods. Our work is also the first-of-its-kind effort to quantitatively evaluate the correlation between user aesthetic preferences and color harmonies of five-color themes, and study such a relation for users with different aesthetic cognition.

CCS Concepts: • **Computing methodologies** → *Image manipulation*; *Machine learning*;

Additional Key Words and Phrases: Color harmony, color theme, machine learning, crowdsourcing, aesthetic cognition

ACM Reference format:
Bailin Yang, Tianxiang Wei, Frederick W. B. Li, Xiaohui Liang, Zhigang Deng, and Yili Fang. 2024. Color Theme Evaluation through User Preference Modeling. *ACM Trans. Appl. Percept.* 21, 3, Article 12 (July 2024), 35 pages.
https://doi.org/10.1145/3665329

1 Introduction

Selecting appropriate color composition is crucial to make art work and graphical design visually appealing, such as painting, images, posters, clothing, and interior home design. Designers or professionals usually make such a

This work was partly supported by the NSFC of China (Grant no. 62172366) and Zhejiang Provincial Natural Science Foundation of China under Grant no. LD24F020003. Zhigang Deng was in part supported by US NSF IIS-2005430.

Authors' Contact Information: Bailin Yang, School of Computer Science and Information Engineering, Zhejiang Gongshang University, Hangzhou 310014, China; e-mail: ybl@mail.zjgsu.edu.cn; Tianxiang Wei, College of Intelligence and Computing, Tianjin University, Tianjin 300072, China; e-mail: tianxiangwei1994@gmail.com; Frederick W. B. Li, Department of Computer Science, University of Durham, Durham, United Kingdom; e-mail: frederick.li@durham.ac.uk; Xiaohui Liang, School of Computer Science and Engineering, Beihang University, Beijing, China; e-mail: liang_xiaohui@buaa.edu.cn; Zhigang Deng (Corresponding author), Department of Computer Science, University of Houston, Houston, TX, USA; e-mail: zdeng4@central.uh.edu; Yili Fang (Corresponding author), School of Computer Science and Information Engineering, Zhejiang Gongshang University, Hangzhou 310014, China; e-mail: fangyl@zjgsu.edu.cn.

ACM 1544-3965/2024/7-ART12
https://doi.org/10.1145/3665329

choice based on their expertise and/or experience, which may be supported by color harmony theory through the application of specific rules and models to achieve aesthetically pleasing color combinations. However, in reality most people lack the artist's expertise or experience; it becomes non-trivial or even challenging for them to properly evaluate color themes.

Some existing websites, such as Adobe Color [1] (formerly known as Adobe Kuler) and COLOURLovers [2], allow users to share color themes with community members and indicate the favor of a color theme by clicking a "Love" or an "Appreciate" button. These services leverage the color wheel's geometric relationships for theme creation, proving especially useful for novices who can either choose from visually compelling color combinations or derive inspiration from existing color themes created by other users to match their aesthetic perception. However, the popularity of these color themes tends to emphasize the spatial relationships between colors, neglecting the complex personal preferences influenced by factors such as users' individual personality traits in their color selections.

O'Donovan et al. [40] collected 10,743 different color themes through **Amazon Mechanical Turk (MTurk)** and trained a prediction model through **Least Absolute Shrinkage and Selection Operator (LASSO)** regression to rate color themes with a 1–5 point scale. Yang et al. [68] trained a model to extract color-pair ratings and then predict color theme preference scores. Despite numerical models' ability to predict the general acceptability of color combinations based on extensive data analysis, these models treat all instances of color theme ratings as if conducted by *the same* anonymous user, failing to capture the unique emotional responses and personalized preferences individuals may have toward certain colors or combinations. The collaborative filtering based method by O'Donovan et al. [41] could infer user variations through matrix factorization. However, they do not work well if color theme ratings in the dataset did not follow a Normal distribution.

Given that previous color theme computation models have overlooked the big literature on human color preferences, we consider the variations in aesthetic cognition among different users and propose a novel color theme evaluation model. To take into account the impact of individual personality traits on individual aesthetic preferences, we created a dataset containing color-themed comments based on the COLOURLovers [2], aimed at gaining deeper insight into users' subjective experiences with colors. We transform comments into user-specific color theme ratings by classifying comments into 1–7 point scales of user preference levels, according to the enclosed affective keywords. We adopt a 7-point scale instead of a five-point one for the sake of reliability as the former offers a wider range of stimuli [12].

We then introduce a kernel probabilistic model based on the output of a trained **back propagation neural network (BPNN)** that can simultaneously infer each user's aesthetics cognition [46] and predict an individual-specific rating for each color theme. Remarkably, our new approach enables us to quantitatively study the correlation between the user preferences and the color harmonies of five-color themes. Our findings can offer fresh insights to designers or relevant professionals for them to decide how to produce artifacts that fit users with different aesthetic preferences. Through various experiments, we demonstrate our method can predict more accurate and more personalized color theme ratings than state-of-the-art methods.

The main contributions of this work include:

- An aesthetic-aware color theme evaluation method;
- A probabilistic model to infer user aesthetic preference; and
- The first-of-its-kind of quantitative evaluation of the correlation between the user aesthetic preferences and the color harmonies of five-color themes.

The remainder of this article is organized as follows. Section 2 presents the recent related work on color modeling and color theme evaluation. Section 3 gives an overview of our proposed framework. Sections 4 and 5 elaborate the technical details of our method, including dataset construction, user aesthetic preference modeling, and color theme evaluation. Section 6 shows a set of experimental results, including comparing our method with state-of-the-art methods through two perspectives of generic evaluation and personalized evaluation, and studying the relationship between user aesthetic preference and color harmony. Finally, we provide concluding remarks and discuss the future work in Section 7.

2 Related Work

2.1 Color Wheels and Palettes

Representing the basic element of visual and graphical communication, colors play an important role in the perception of visual design [65]. The origin of color harmony can be traced back to the Newton's color theory, which emphasizes that multiple colors in adjacent areas can produce visually pleasing effects. Color harmony models are divided into geometric, numerical, and contingent models, with the color wheel being a quintessential representation of the geometric models [8, 36]. Through different positions of colors in the color wheel, a harmonious color scheme can be generated, such as complementary, analogous, and triadic. Itten [20] designed a 12-color wheel, namely the base color wheel, assuming the colors uniformly distributed on the color wheel are harmonious. Based on this concept, Matsuda [33] derived a set of 8 hue templates according to fashion questionnaires and color themes used in fashion industry, which have been widely used for many applications, including custom color themes [1, 2, 16, 56], aesthetics assessment [30, 31, 39, 40], and image optimization [13, 25, 47, 62]. Readers may refer to [66] for a comprehensive review on different color harmony models.

A color palette comprises a number of different solid color swatches, being commonly used in design applications. Online coloring websites including Adobe Color [1] and COLOURLovers [2] allow users to quickly pick favorite colors and create new palettes. Color wheels are used to visualize arbitrary color palettes in a more contextual way, in order to infer meaningful relationships between colors. By contrast, a given color palette can be shown in an order that obfuscates any harmonic relationship, making it less intuitive. As a result, color palettes have attracted a lot of attentions in research communities, such as extracting quantitative palettes from images [9, 27, 29], color transfer [18], and color palette based recoloring [9, 10, 27, 61, 70]. Cao et al. [6] presented a probabilistic model to learn color palettes from a collection of artworks by summarizing their color uses. Shugrina et al. [57] designed a novel interactive interface for generating color palettes that are more consistent with the color theory, which can intuitively relate multiple color design tasks.

2.2 Color Preference

Color preference refers to an individual's degree of liking specific colors, and it is considered as a subarea of color psychology [34]. Color preference is influenced by various factors such as age [63], mood [59], education [55], personal preferences [45], season [53], and culture [69]. For example, researchers found that western individuals typically prefer cool colors over warm colors [19, 28]. Saito suggested that white is the favorite color among East Asian individuals, associated with concepts of cleanliness, purity, harmony, freshness, beauty, happiness, gentleness, and nature [50]. Palmer and Schloss studied aesthetic preferences in cognitive science, introducing the **ecological valence theory (EVT)** for color preference[45, 46]. EVT posits that an individual's preference for a given color at a particular time depends on (a) the concepts associated with that color (differential object-association hypothesis), (b) the relative preference for concepts associated with that color (differential valence hypothesis), and (c) the degree to which different concepts are activated in an individual's mind (differential activation hypothesis).

Building upon the EVT, researchers have ventured into modeling color preferences. Schloss et al. [52] developed a color inference framework that interconnects color-related concepts to determine preferences for specific colors. Additionally, researchers have explored various metrics of the color space to construct models that predict preferences for colors that users have not previously assessed [19, 51]. However, these methods are limited to predicting preferences for individual colors. In real world scenarios, colors often appear in combinations of multiple colors, and the combination of colors usually involves color harmony, which is a key factor influencing color preferences. Different hues, areas, and arrangements of colors are linked to diverse physiological and psychological perceptions [43, 46]. Palettes serve as intuitive visual representations of these elements. Interactive color recommendation systems based on community-generated palettes have been developed, facilitating users in instantaneously selecting harmonious colors [3, 58]. Colorgorical [16], a palette generation tool designed for information visualization, takes into account pairing preferences between color pairs and is employed for predicting higher-order color combinations.

2.3 Color Theme Evaluation Models

Two primary research streams address the evaluation of color themes. In the field of psychology, various models have been proposed to predict harmonious color combinations. These models were designed to explore the factors that satisfy the harmony of color combinations through psychological experiment data [42, 43, 62]. On the other hand, researchers construct computational models to evaluate color themes based on existing color rating datasets. For instance, O'Donovan et al. [40] conducted experiments to study color ratings from large online datasets for predicting color theme ratings, which overcomes the robustness and generalization problems of previous studies due to insufficient data. Kita and Miyata [22] proposed a model to add a color to any position of a known color theme while maintaining the harmony of a newly created color theme. Compared to previous methods, their model could rate any number of color palettes, although their results are less accurate than [40]. Based on the key observation that color pairs play an influential role to a color theme, Yang et al. [68] proposed a two-layer, color-pairs based model to first predict the scores of color pairs in a color theme, and then statistically combine the scores of the enclosed color pairs to generate the final rating of the color theme. These existing models however simply treat all instances of color theme ratings as if they were done by a single anonymous user, without considering the aesthetic cognition differences among individuals. O'Donovan et al. [41] randomly extracted 13,343 color themes evaluated by 1,137 participants from the Adobe Kuler website, and proposed a collaborative filtering based approach to capture the user preferences of color combinations. However, this method overlooks the impact of personality traits on color preferences, which are considered crucial for distinguishing between color harmony and individual aesthetic preferences [44].

Inspired by the findings of Schloss et al. [54], which demonstrated a high correlation between user preferences and color harmonies of color pairs (the Pearson correlation coefficient $r = +0.79$), we acknowledge a crucial observation: individuals exhibit different aesthetic cognition and preferences when evaluating the same color theme. This allows us to develop a more accurate and personalized model for predicting color theme ratings. It also enables us to quantitatively evaluate the correlation between user preferences and color harmonies of color themes under different user aesthetic cognition.

3 Our Approach

We now give an overview of our framework, describing the major steps involved. First, based on the data from COLOURLovers, we construct a dataset by collecting the color themes C_n, user identifications m, user-specific color theme ratings Z_{mn}, and the set of color themes C_{mn} annotated by each user. Second, using the collected dataset as input, we train our BPNN model using a set of features extracted from color themes to infer the average rating β_n. After that, we feed the output into a probabilistic density estimation model to predict color-theme-independent user aesthetic preferences α_m. Finally, our framework can produce user-specific color theme rating Y_{mn}. In addition, we can apply majority voting to generate a single rating for each color theme by counting the ratings on each distinct color theme. Figure 1 illustrates how the major components of our framework work together.

We define the following three terms for the sake of clarity.

- *True Ratings* of a color theme: They refer to the ground-truth annotator-specific preference ratings on a color theme. In this work, they are directly obtained from the COLOURLover dataset, as described in Section 4.

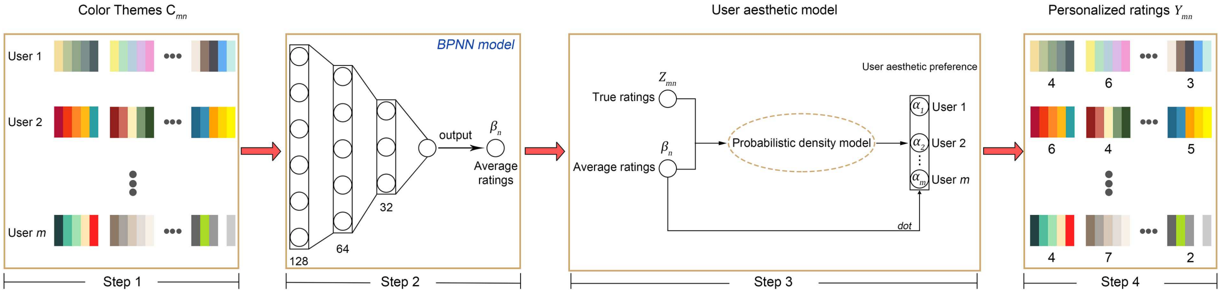


Fig. 1. The pipeline of our color theme evaluation framework consists of the following major steps. *Step 1*: Extend color themes such that each user m has a corresponding color theme C_{mn} associated with the rating Z_{mn} for color theme n. *Step 2:* Extract features from the color themes, and generate the average ratings β_n through a trained BPNN model. *Step 3:* Predict the aesthetic preference α_m corresponding to each user through a probabilistic density model. *Step 4:* Generate final individual-specific ratings Y_{mn} on color themes based on the aesthetic preference of each user.

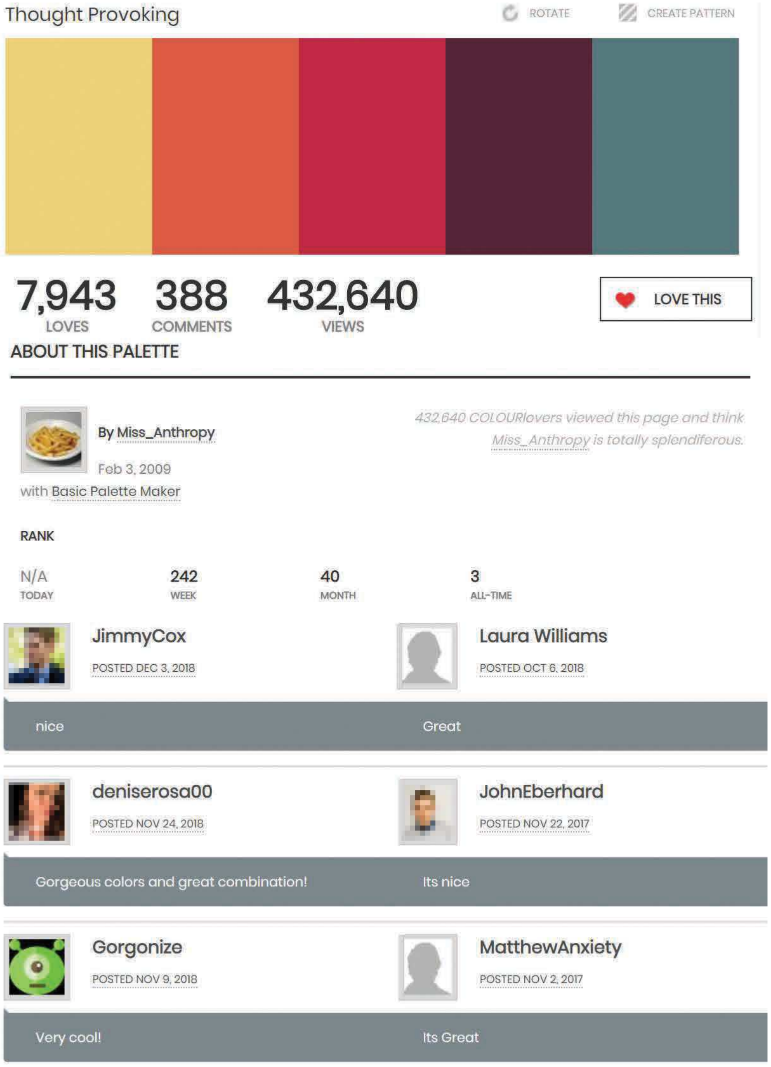
- *Average Rating* of a color theme: It refers to the generally accepted rating of a color theme by different annotators. The "average" denotes that the rating of a color theme remains uninfluenced by individual annotators' preferences. Unlike true ratings, each color theme has only one average rating.
- *Predicted Rating* of a color theme: It refers to an output or predicted rating by our computational model or other algorithms, given an input color theme.

4 Dataset Construction

To the best of our knowledge, datasets for studying user preference in color theme are limited. The most prominent datasets [40] were derived from Adobe Kuler [1], COLOURLovers [2], and MTurk. The one derived from COLOURLovers plays a dominant role in some studies [40] due to its large data size. These datasets were intrinsically biased due to the anonymous single user rating issue. Remarkably, our proposed dataset comprises user-specific color theme ratings. As our method also utilizes the dataset from COLOURLovers, experimental comparisons between our method and existing methods should be sufficiently fair.

Different from existing works, our model takes user-specific color theme ratings (or called true ratings, as defined above) as input. Specifically, our dataset was collected from the COLOURLovers, a website maintaining over one million different 2–5 color themes defined by users. The website also contained over 4,670,225 user accounts as of 1 January 2020. The website did not support users to directly rate color themes, but allowed users to indicate color preference by clicking the "heartening" button. Unfortunately, no quantified user-specific ratings or preferences can be retrieved. To solve this problem, we alternatively collected users' online comments on different color themes, as shown in Figure 2. We have collected 256,624 comments (mostly commented by 36,293 different users) together with user information for 26,470 different five-color themes.

In the field of sentiment analysis, researchers use specific keywords as a corpus of sentiment intensities, and further use the corpus to automatically classify sentences into sentiment levels [35, 49]. Inspired by the above research, in this work, we also treat the collected user comments as an unmarked corpus. We select the scoring method of Likert scale [14, 21] in the field of psychology and classify user comments into seven different levels, where one represents the least favored level and seven represents the most favored level, based on the Hourglass of Emotions model [60]. In this model, both positive and negative keywords are classified into three polarity categories. Each polarity category contains four basic emotional dimensions, and the fusion of these basic emotions can generate more compound emotions. Figure 3 shows some emotion classification examples. After judging the emotional polarity of the keywords, we analyze the polarity of the keywords through SenticNet [5] to



Thought Provoking

7,943 LOVES
388 COMMENTS
432,640 VIEWS

ABOUT THIS PALETTE

By Miss_Anthropy
Feb 3, 2009
with Basic Palette Maker

432,640 COLOURlovers viewed this page and think Miss_Anthropy is totally splendiferous.

RANK

N/A	242	40	3
TODAY	WEEK	MONTH	ALL-TIME

JimmyCox
POSTED DEC 3, 2018
nice

Laura Williams
POSTED OCT 6, 2018
Great

deniserosa00
POSTED NOV 24, 2018
Gorgeous colors and great combination!

JohnEberhard
POSTED NOV 22, 2017
Its nice

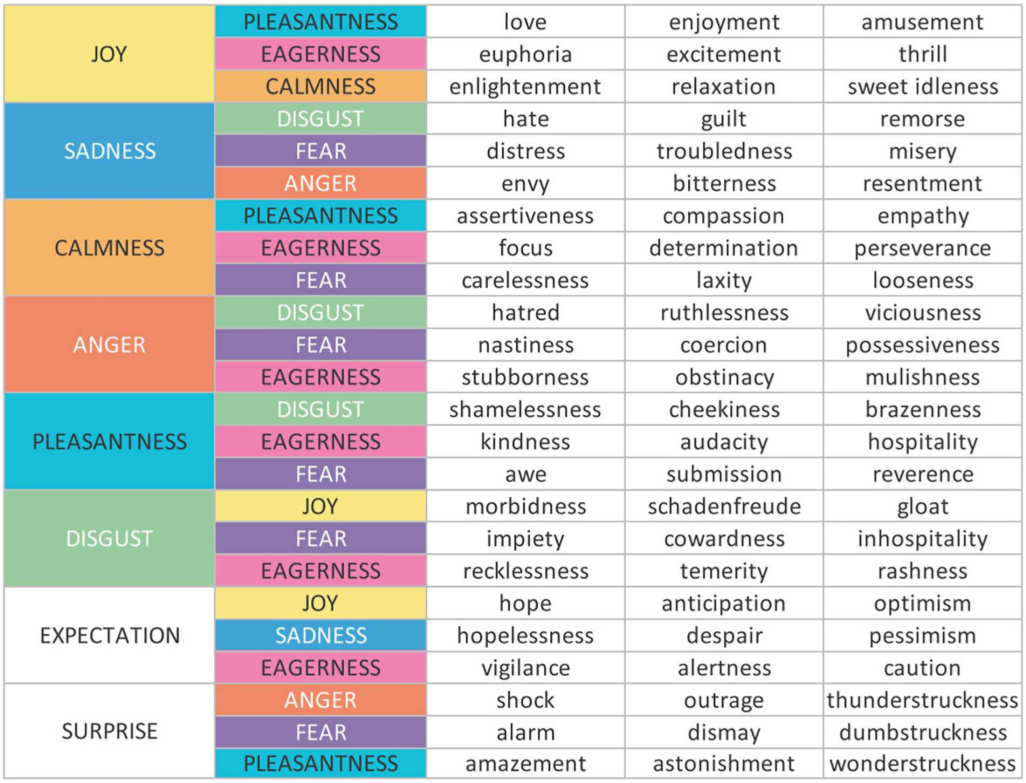
Gorgonize
POSTED NOV 9, 2018
Very cool!

MatthewAnxiety
POSTED NOV 2, 2017
Its Great

Fig. 2. Top: A five-color theme created by MissAnthropy, which is called "thought provoking," on 1 October 2019. Bottom: With the comment entries of this color theme, we can retrieve identifiable users and their comments, quantifying user-specific color theme rating accordingly.

obtain the final polarity of each comment. By determining positive or negative vocabulary keywords contained in the comments, we then quantify the corresponding degree of user preference.

We chose keywords that can indicate user preferences distinctively and explicitly. For comments without any keywords, we quantified their degree of user preference as four. To quantify the degree of preference for emojis, we classified positive ones based on smile faces and negative ones based on crying faces to have the degrees of preference of five and three, respectively. We removed blank comments and their corresponding users to reduce noise in our dataset. Each color theme is identified by a unique ID. To obtain the IDs of the users (i.e., annotators), we followed each color theme, sorting the annotators by their usernames and enumerating them accordingly. Note that we only build a personalized model from the perspectives of both users' visual perception and color



JOY	PLEASANTNESS	love	enjoyment	amusement
	EAGERNESS	euphoria	excitement	thrill
	CALMNESS	enlightenment	relaxation	sweet idleness
SADNESS	DISGUST	hate	guilt	remorse
	FEAR	distress	troubledness	misery
	ANGER	envy	bitterness	resentment
CALMNESS	PLEASANTNESS	assertiveness	compassion	empathy
	EAGERNESS	focus	determination	perseverance
	FEAR	carelessness	laxity	looseness
ANGER	DISGUST	hatred	ruthlessness	viciousness
	FEAR	nastiness	coercion	possessiveness
	EAGERNESS	stubborness	obstinacy	mulishness
PLEASANTNESS	DISGUST	shamelessness	cheekiness	brazenness
	EAGERNESS	kindness	audacity	hospitality
	FEAR	awe	submission	reverence
DISGUST	JOY	morbidness	schadenfreude	gloat
	FEAR	impiety	cowardness	inhospitality
	EAGERNESS	recklessness	temerity	rashness
EXPECTATION	JOY	hope	anticipation	optimism
	SADNESS	hopelessness	despair	pessimism
	EAGERNESS	vigilance	alertness	caution
SURPRISE	ANGER	shock	outrage	thunderstruckness
	FEAR	alarm	dismay	dumbstruckness
	PLEASANTNESS	amazement	astonishment	wonderstruckness

Fig. 3. Example of emotion classification words by Susanto et al. [60]. The first and second columns are basic emotion words. The compound sentiment words in columns 3–5 are obtained by fuzing different basic emotion words.

theme features. Although some personal user information exists in COLOURLovers for potentially enriching our dataset, such as age, gender, nationality, occupation, and these data lack completeness. For example, among a total of 36,293 users, 36% of users lack age and gender, and only 8% of users disclose all the above personal information.

Color preference analysis: Although the color preferences of participants in the dataset were derived from assessments of color themes, the specific color attributes associated with individual color differences remain unclear. To investigate this, we employed **multidimensional scaling (MDS)** [7] to analyze clusters of participants' preferences. We selected participants who evaluated more than 50 color themes, culminating in a sample size of 1,220 individuals. For the purpose of correlation analysis, we identified each participant's preferred color theme, which was denoted by a top score of seven, and considered this as a reflection of their highest color preference. Our MDS model incorporated various perceptual color factors, such as the saturation, brightness, hue, color distance, and color contrast entropy of the color themes. The results are shown in Figure 4, presenting each participant's color preference through individual scatter points. We use colors to differentiate scatterplot densities. Blue represents the dense areas, which are predominantly clustered around the origin, indicating that these participants have similar color preferences. In contrast, red denotes the sparse areas of the plot, which are more dispersed and do not form significant clusters, suggesting that these participants have more diverse and scattered color preferences without clear trends or easily distinguishable groups.

To further analyze the preference differences between these two participant groups, we analyzed the distribution of color attributes for these two groups separately. As shown in Figure 5, we computed the color distance, contrast,

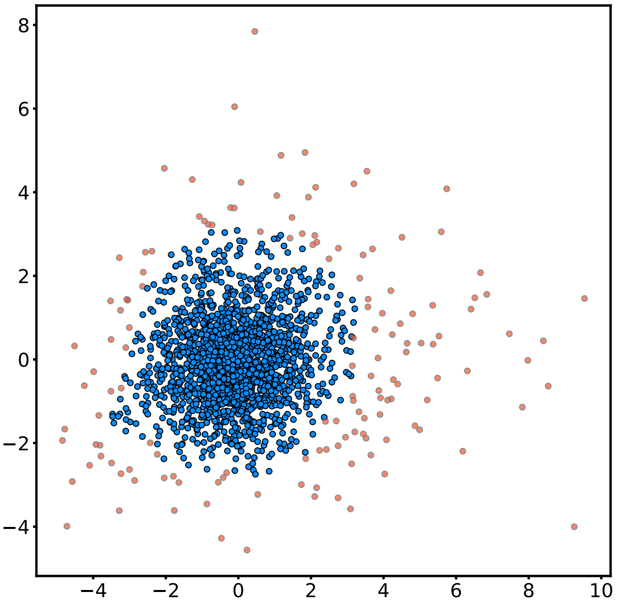


Fig. 4. Color preferences of participants were analyzed using MDS. Blue scatter points indicate dense clusters, representing similar individual color preferences. Red scatter points indicate sparse clusters, representing unique individual color preferences.

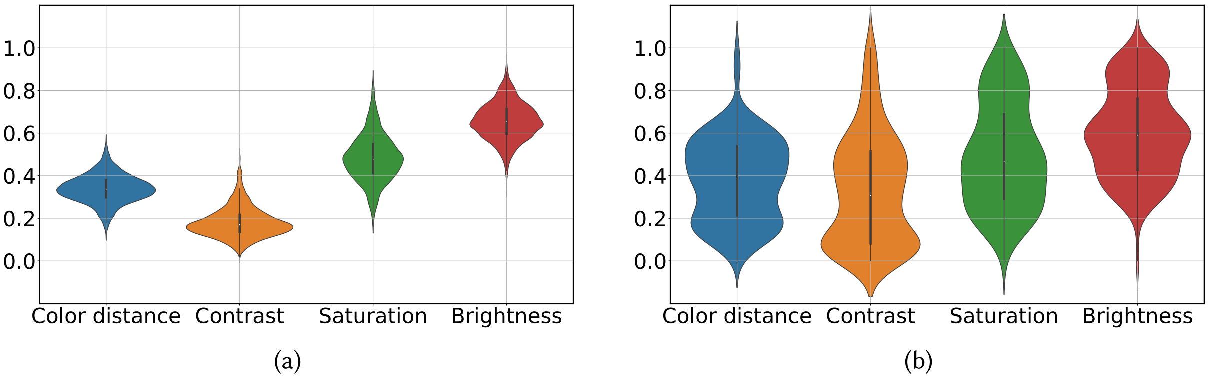


Fig. 5. The color distribution of color themes preferred by participants was analyzed from different color attributes. (a) A group of participants with similar preferences. (b) A group of participants with greatly different preferences.

saturation, and brightness of the color themes favored by diverse participants and normalized these attributes for representation. Figure 5(a) depicts the distribution of color attributes among participants with similar preferences. We can observe that these color attributes approximately conform to a Gaussian distribution, with each attribute exhibiting a narrow distribution range. This indicates a collective predilection for color themes within a specific attribute range among these participants. In contrast, the distribution of color attributes among participants with unique color preferences, as illustrated in Figure 5(b), although maintaining a similar median, spans a wider range and exhibits more irregularity. This implies a more varied array of favored colors within this subset. The elongated tails of the distribution underscore a greater dispersion in color preferences. For example, a segment of these

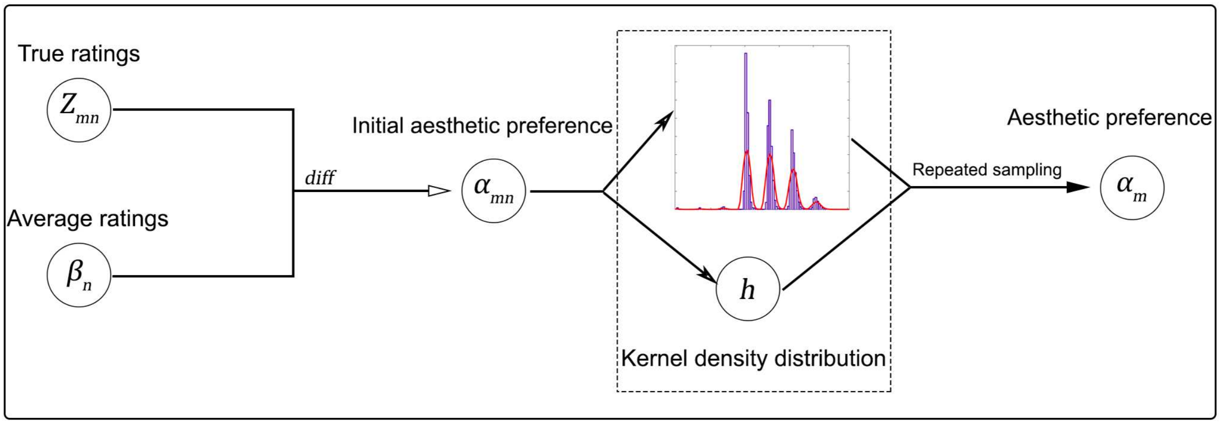


Fig. 6. User aesthetic preference α_m obtained via probability density estimation.

participants favored color themes with a color distance greater than 0.8, marking a substantial divergence from the preferences of their counterparts. This analysis further suggests that participants with similar preferences tend to be attracted to similar color themes, whereas those with unique preferences favor a diversity of color themes. Although perceptual color attributes are inextricably linked to color preferences, the individual color preferences exhibit a high degree of complexity, making them challenging to capture with simple statistical models. This necessitates the crafting of bespoke preference models to prognosticate the color preferences of participants with greater accuracy.

5 User Aesthetic Model

Schloss and Palmer [54] proposed a high correlation between user preference and color harmony. Concurrently, Yang et al. [68] indicated that color harmony might be nonlinearly correlated with certain low-level features of color themes. This inspires us to extend BPNN to learn such a relationship, aiming to predict both color theme ratings and the user preference to judge aesthetics. As illustrated in Figure 1, we can infer the average color theme ratings through the BPNN network, and predict the users' initial (i.e., color theme specific) aesthetic preferences. Such predicted aesthetic preferences are *color theme specific* since the color theme ratings in our dataset have the same nature. To generalize the prediction of the user preference on unseen color themes, we need to obtain the *color-theme-independent* user aesthetic preference. This is analogous to solving the worker modeling problem in crowd-sourcing research [23]. Although we are aware that the collaborative filtering based approach [41] can infer user-specific features to a certain extent, our experimental results show that such an approach is not effective enough. To tackle this, we devise a probabilistic density estimation model as illustrated in Figure 6.

5.1 Model Training

To employ low-level color features for BPNN model training, we follow the work in Yang et al. [68] to extract features from a color theme, including the colors enclosed in the color theme, mean, standard deviation, median, max, min, mode, color moment, max-minus-min across a single channel in each color space (i.e., RGB, LAB, HSV, and LCH), and the Euclidean distances between the adjacent colors in the color theme. These features reflect the impact of the specific color features such as color difference, color sequence, hue, saturation, and so on, on user color preferences. Note that, different from the work of Yang et al. [68], we do not use a two-layer maximum likelihood estimation to produce coarse color-pair ratings as features. We obtain a total of 184 features from each color theme as the input to our BPNN model.

For network design, the input layer of our BPNN model consists of 184 neurons for the input feature vector while the output layer consists of only one neuron for its prediction. We choose three hidden layers to train the color features of the input layer and determine the number of nodes in each hidden layer through multiple experiments. The three hidden layers have 128, 64, and 32 neurons, respectively. This design is experimentally chosen to balance the tradeoff between training time and regression accuracy, while having a sufficient capability to capture the nonlinear relations in different color themes. After the output of the third hidden layer, only a scalar value is needed as the prediction outcome (i.e., the average rating of the input color theme), which is used as one of the inputs to the subsequent probabilistic density estimation model.

We use a *sigmoid* function as the activation function in the three hidden layers. Between the third hidden layer and the output layer, we use the *purelin* linear transfer function (i.e., $y = x$) as the activation function. After that, we then feed the result of the output layer, as well as the true ratings of the color themes, into a probabilistic density estimation model to infer the color-theme-independent user aesthetic preference.

To facilitate model training, we set the learning rate $\eta = 0.00075$ to balance prediction accuracy and training time. Since if the learning rate is too large, the loss function may not converge after many iterations. On the other hand, a small learning rate may not achieve the optimal convergence. The learning rate is set to decrease over iterations. This ensures that our model remains stable with more iterations. The number of iterations is important to obtain the optimal prediction result. We experimentally set the maximum number of iterations to 500. If the losses resulted from three consecutive iterations were approximately the same (i.e., the difference is below a threshold), the training process will then be stopped.

5.2 Aesthetic Preference and Color Theme Evaluation

A novel contribution of this work is the capability of predicting the user's preference to judge aesthetics on color themes. To enable this, as depicted in Figure 6, our BPNN model accepts true ratings and average ratings on color themes as the inputs to generate *initial aesthetic preference values*. Since the values are color theme specific, they cannot be directly used to represent the generalized aesthetic preferences of users. Therefore, we introduce a probabilistic density estimation model to generate *color-theme-independent, user aesthetic preferences*, based on the initial user aesthetic preferences.

5.2.1 User Aesthetic Preference Modeling. Whitehill et al. [67] proposed an algorithm to control the quality of crowdsourced data by applying a generative model to infer the expertise of different users. Inspired by this work, we collected color theme ratings by different users and improve color theme evaluation by modeling the aesthetic preferences of the users. Since aesthetic is a subjective preference. Therefore, "aesthetic preference" does not mean that something can be done correctly or incorrectly. It is similar to the user's differences in aesthetic cognition. Unlike [67], we build a kernel-based probabilistic density model to simulate and thus generate color theme ratings for different users. In addition, we have explored alternatives to our framework design, including the replacement of **kernel density model (KDM)** with a neural network decoder, as shown in Section 6. We found that the configuration of our current framework design (refer to Figure 1) performs the best.

To explain our probabilistic density estimation model, let N be the number of color themes and M be the total number of users (i.e., annotators) involved. User aesthetic preferences are represented by $\boldsymbol{\alpha} = \{\alpha_1, ..., \alpha_m, ..., \alpha_M\} \in [0, 1]$, where m denotes the mth user. The average ratings of color themes predicted by our BPNN model in Section 5.1 are normalized and represented by $\boldsymbol{\beta'} = \{\beta_1, ..., \beta_n, ..., \beta_N\} \in [0, 1]$. The true rating of a color theme C_n given by a user m is represented by Z_{mn}. The total number of Z_{mn} is S, where $S \leq M \times N$ since not all of the users rated all the color themes.

In this work, $\boldsymbol{\alpha}$ is the key factor that affects individual-specific ratings on color themes. Specifically, $\alpha_m = \frac{1}{2}$ indicates the user m possesses the professional artist's aesthetic cognition, being able to give an accurate evaluation on different color themes; $\alpha_m < \frac{1}{2}$ indicates that the user m generally gives lower ratings to color themes, compared to the average user; and $\alpha_m > \frac{1}{2}$ indicates that the user m generally gives higher ratings to color themes, compared to the average user, and thus the ratings given by the user m are typically higher than the average ratings of color themes. As extreme cases, $\alpha_m = 0$ or 1 indicates that the user m is extremely underestimates or overestimates the aesthetic quality of color themes. In our model, the *predicted* individual-specific rating of the user m on the color theme C_n is denoted as Y_{mn}; it is composed of the aesthetic preference α_m of the user m and the average rating β_n predicted by the BPNN model: $Y_{mn} = 2\alpha_m.\beta_n$; if $2\alpha_m.\beta_n > 1$, then we force $Y_{mn} = 1$.

In this way, $Y \in \mathbb{R}^{S\times 1}$ represents all the predicted user-specific color theme ratings generated by our model. We multiply a coefficient 2.0 in the computation of Y_{mn} so that the average color theme rating is the same as the predicted result when $\alpha = \frac{1}{2}$. We then formulate the learning of user aesthetic preference as a regression problem. We use the mean squared error as a multi-personality regression loss function and calculate the gradients of model parameters as follows:

$$L = \frac{1}{S}\sum_{\substack{m=1,\\ n=1}}^{S}(Z_{mn} - Y_{mn})^2 = \frac{1}{S}\sum_{\substack{m=1,\\ n=1}}^{S}(Z_{mn} - 2\alpha_m.\beta_n)^2. \tag{1}$$

Obviously, $\{\alpha_m\}$ are the parameters to be solved, and solving them is similar to the solving of the worker modeling problem in the area of crowdsourcing research [23]. In our approach, we predict users' color-theme-independent aesthetic preferences by exploiting the probabilistic distribution of all instances of users' color-theme-specific aesthetic preferences.

Given that our BPNN model can predict $\{\beta_n\}$ and true ratings $\{Z_{mn}\}$ were obtained in our dataset, we can evaluate the color-theme-specific aesthetic preference α'_{mn} of user m as:

$$\alpha'_{mn} = \frac{Z_{mn} - \beta_n}{2} + \frac{1}{2}, \tag{2}$$

where α'_{mn} is referred to as an *initial aesthetic preference* value. Note that $Z_{mn} - \beta_n$ and α'_{mn} are linearly correlated. When $Z_{mn} = \beta_n$, the initial aesthetic preference value $\alpha'_{mn} = \frac{1}{2}$.

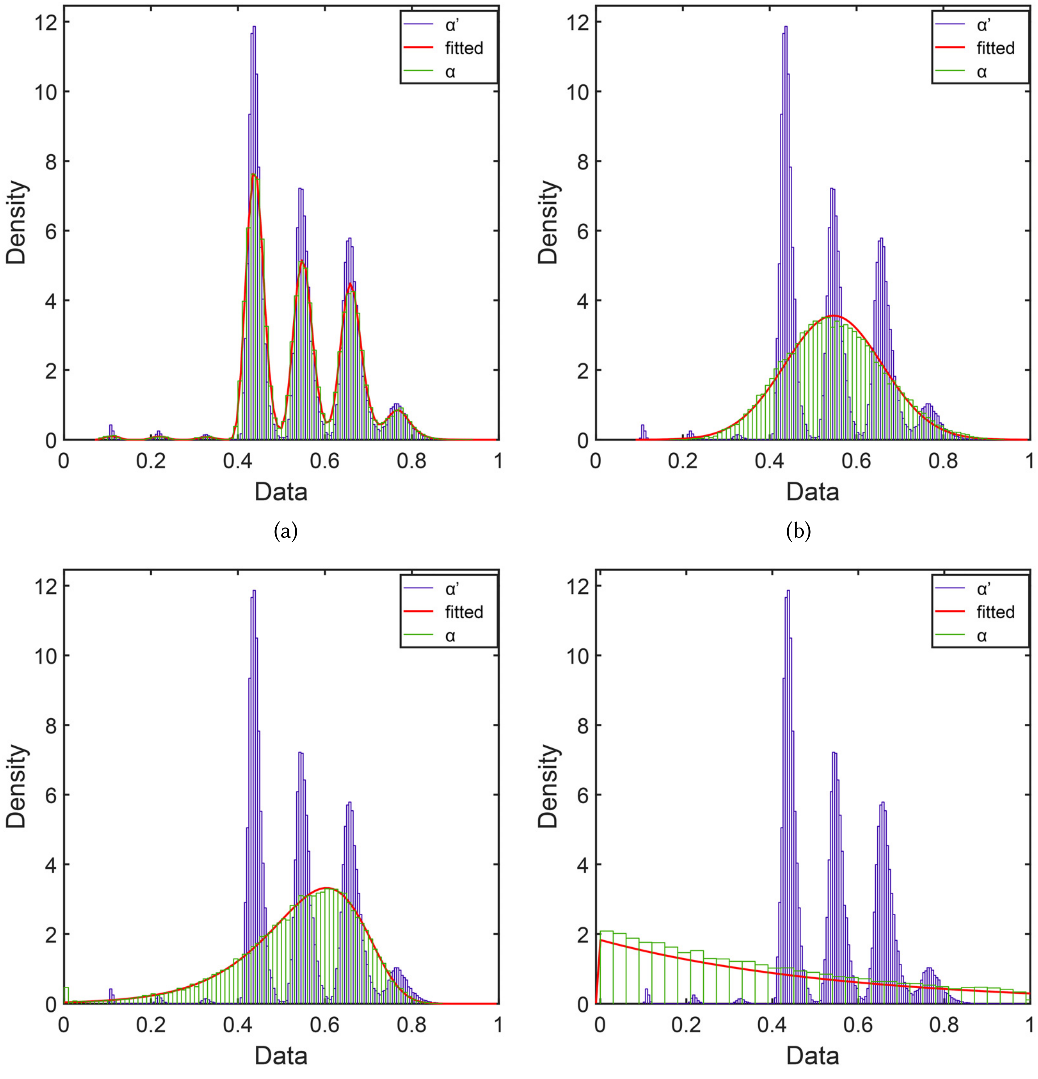
5.2.2 Color-Theme-Independent, User Aesthetic Preference Generation. In order to predict user preferences for unseen color themes, we fit the aesthetic preference of users through probability distributions. To analyze the distribution of $\{\alpha'_{mn}\}$, we experimented with some different types of widely-used probabilistic distribution models as depicted in Figure 7. We found that the distribution of $\{\alpha'_{mn}\}$ can well fit into the kernel distribution as shown in Figure 7(a), while the other types of distributions can hardly fit the distribution of $\{\alpha'_{mn}\}$. We therefore derive color theme independent, aesthetic preference values of the users through the fitting of kernel distributions.

Since we independently solve the color-theme-independent, aesthetic preference value for each user, without the loss of generality, we describe how we compute the α_m of a specific user m. Assuming from the previous steps we have obtained l_m initial aesthetic preference values of the user m, $\{\alpha'_{m1}, ..., \alpha'_{ml_m}\}$, now we model the probabilistic distribution of α_m by fitting a kernel distribution to $\{\alpha'_{m1}, ..., \alpha'_{ml_m}\}$.

Specifically, first, we obtain the bandwidth parameter h and evaluate the probability density function $p(\alpha_m|K, h)$ through the following kernel density estimator:

$$p(\alpha_m|K, h) = \frac{1}{l_m h}\sum_{j=1}^{l_m} K(\alpha_m, \alpha'_{mj}, h)\left(\frac{\alpha_m - \alpha'_{mj}}{h}\right), \tag{3}$$

$$h = \left(\frac{4}{3l_m}\right)^{\frac{1}{5}}\sigma, \tag{4}$$



(a)

(b)

(c)

(d)

Fig. 7. Probability density histogram comparisons between the initial aesthetic preference values ($\boldsymbol{\alpha'}$) and the aesthetic preference values ($\boldsymbol{\alpha}$) generated by (a) Kernel distribution, (b) Gaussian distribution, (c) Extreme value distribution, and (d) Exponential distribution, respectively. For each type of distribution, the red curve (fitted) shows its fitted probability density curve.

$$K(\alpha_m, \alpha'_{mj}, h) = \frac{3}{4}\left(1 - \frac{(\alpha_m - \alpha'_{mj})^2}{h^2}\right), \tag{5}$$

where σ is the standard deviation of $\{\alpha'_{m1}, ..., \alpha'_{ml_m}\}$, and h is the bandwidth parameter. We adopt the bandwidth estimator proposed by Dehnad [15]. In addition, K is a scaled kernel function. Based on our experiments, we empirically choose the Epanechnikov kernel for estimation.

Second, we use the method in Baszczynska [4] to change the probability density function $p(\alpha_m|K, h)$ into a cumulative distribution function $F(\alpha_m|K, h)$ as follows:

$$\begin{aligned} F(\alpha_m|K,h) &= \frac{1}{l_m}\sum_{j=1}^{l_m}\int_{-\infty}^{\frac{\alpha_m-\alpha'_{mj}}{h}} K(\alpha_m,\alpha'_{mj},h)d\left(\frac{\alpha_m-\alpha'_{mj}}{h}\right) \\ &= \frac{1}{l_m}\sum_{j=1}^{l_m} W(\alpha_m,\alpha'_{mj},h). \end{aligned} \tag{6}$$

$$W(\alpha_m,\alpha'_{mj},h) = -\frac{1}{4}\frac{(\alpha_m-\alpha'_{mj})^3}{h^3} + \frac{3}{4}\frac{(\alpha_m-\alpha'_{mj})}{h} + \frac{1}{2}. \tag{7}$$

In the above Equations (6) and (7), $W(\alpha_m, \alpha'_{mj}, h)$ is a kernel function of the cumulative distribution in the case of the Epanechnikov kernel.

Finally, we use the above cumulative distribution function to inversely compute the aesthetic preference α_m satisfying the Kernel distribution. We first generate a group of random variables $\{u_i\}$ that satisfy the uniform distribution [0,1], and then we generate their corresponding color theme independent, aesthetic preference values $\{\alpha^*_{mi}\}$ by applying inverse transform sampling to the cumulative distribution function as follows:

$$\alpha^*_{mi} = F^{-1}(u_i) = \frac{1}{l_m}\sum_{j=1}^{l_m} W^{-1}(u_i,\alpha'_{mj}). \tag{8}$$

$$\begin{aligned} W^{-1}(u,\alpha'_{mj}) &= h(2\sqrt{(1-u)u}-2u+1)^{\frac{1}{3}} \\ &\quad + h(2\sqrt{(1-u)u}-2u+1)^{-\frac{1}{3}} + \alpha'_{mj}. \end{aligned} \tag{9}$$

Finally, we average the computed $\{\alpha^*_{mi}\}$ to obtain the final α_m as the color-theme-independent aesthetic preference value of the user m. In our experiments, we sampled u 1000 times and then performed the averaging operation.

5.2.3 Color Theme Rating Generation. After the above α_m is obtained, we can generate the aesthetic-aware rating on a color theme C_i by the user m as follows: (1) We first use our BPNN model to predict its average rating β_i; (2) then, we further compute its aesthetic-aware rating by the user m as $Y_{mi} = 2\alpha_m{\cdot}\beta_i$, and if $2\alpha_m{\cdot}\beta_i$ is larger than 1.0, then we truncate Y_{mi} to 1; (3) finally, we linearly scale Y_{mi} from its [0, 1] range to the seven-point scale. Figure 8 illustrates how aesthetic-aware color theme ratings are computed. In the figure, annotator #1 evaluates all three color themes. Annotator-specific color theme ratings Y_{11}, Y_{12}, and Y_{13} are then evaluated based on the computed aesthetic preference of annotator #1, α_1, and the corresponding average color theme ratings, β_1, β_2, and β_3, outputted by the BPNN model. Note that annotator #2 only rates two color themes instead of all of them.

After obtaining the aesthetic-aware color ratings of different users, $\{Y_{mn}\}$, we can calculate the losses according to Z_{mn} and Y_{mn} for back-propagation iterations and updating the weights. In this way, we obtain an updated kernel density bandwidth h, an updated probability density function $p(\alpha_m|K, h)$, and an updated cumulative distribution function $F(\alpha_m|K, h)$. Finally, $\{\alpha_m\}$ will be updated accordingly. This iterative process continues until convergence (i.e., the loss difference between two consecutive iterations is smaller than a threshold or a maximum number of iterations are achieved). In each iteration, $\{\alpha_m\}$ and $\{\beta_n\}$ are trained simultaneously.

6 Experiment Results and Evaluations

We present experimental results on test color themes and personalized color theme datasets using our approach. We then present a set of quantitative tests to validate the effectiveness and robustness of our method, and some user study results.

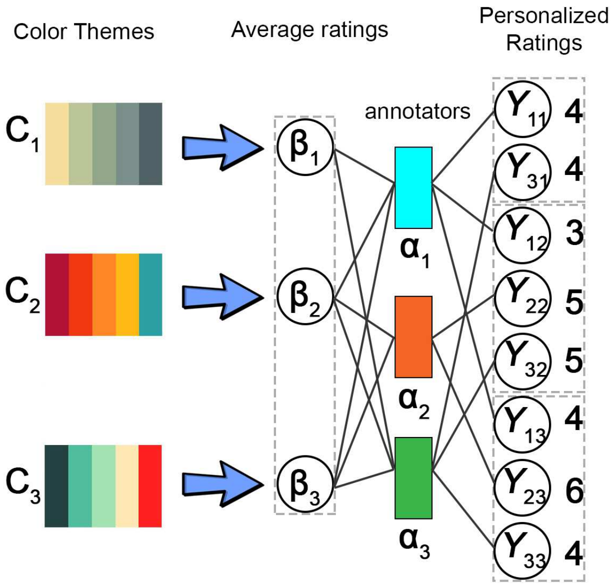


Fig. 8. An illustration of the process of generating aesthetic-aware color theme ratings Y_{mn}. The example involves three color themes C_1, C_2, C_3 and three annotators $\alpha_1, \alpha_2, \alpha_3$.

6.1 Color Theme Assessment

To evaluate the effectiveness of our method on predicting the generic (in contrast to personalized) rating of color themes, we tested our method on both the MTurk dataset [41] and our collected dataset (as described in Section 6). Specifically, the MTurk dataset is the only publicly available dataset with personalized aesthetic ratings for color themes. This dataset comprises 13,343 five-color themes downloaded from the Adobe Kuler website [1], which were rated by a total of 1,137 annotators. From both the datasets, we randomly selected 60% of the data for training, 20% for validation, and the remaining 20% for test. With our method, we learn the aesthetic preference of each annotator and then predict user preference levels of the color themes in the test set for each annotator. Finally, we use majority voting to obtain the final rating of each color theme in the test set. Figure 9 shows the predictions of some testing color themes by our method, compared to the ground truth (i.e., true labels). As shown in this figure, our method can generate reasonably accurate ratings for test color themes with small deviations from the ground truth.

Accuracy: To evaluate the accuracy of our method, we used two widely-used regression quantitative metrics: **root mean squared error (RMSE)** and **mean absolute error (MAE)**. We compared our approach with state-of-the-art color theme evaluation methods, including the color pairs based method [68], the LASSO regression based method [40], and the collaborative filtering based method [41]. To ensure a fair comparison, we trained all the methods using our dataset (refer to Section 6) and applied a seven-point scale for color theme rating across the board. As shown in Table 1, in terms of RMSE (or MAE), the accuracy of our method is about 30% (or 18.0%) more than that of the color pair based method [68] and about 46.7% (or 34.1%) more than that of the LASSO regression based method [40]. This proves that taking into account user-specific aesthetic preferences can significantly improve the accuracy of color theme rating prediction. Our work also outperformed the collaborative filtering based method [41], in particular, by about 15.5% in terms of RMSE.

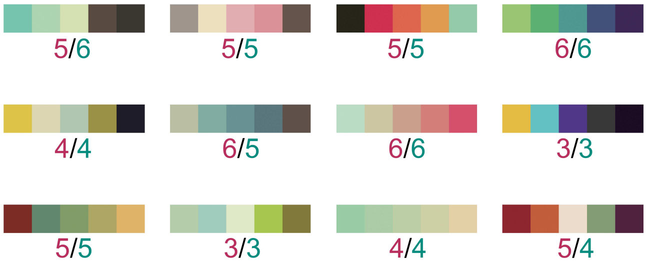


Fig. 9. Comparisons between the predicted preference ratings by our method (red) and the ground truth (dark cyan) for some color themes in our test set.

Table 1. Color Theme Evaluation Performances of Different Methods Using Both Regression Quantitative Metrics and Categorical Quantitative Metrics, Based on Our Collected Dataset (Described in Section 6)

Metric / Model	RMSE ↓	MAE ↓	Acc(%) ↑	PLCC ↑
LASSO regression based [40]	1.2484	0.8264	53.3	0.1218
Color pairs based [68]	0.9464	0.6642	63.7	0.2698
Collaborative filtering [41]	0.7872	0.5862	74.9	0.3937
Alternative combination (BPNN + decoder)	0.9779	0.7147	59.2	0.2454
Alternative combination (InceptionV3 + KDM)	0.6944	0.5727	77.2	0.4002
Our method	**0.6650**	**0.5445**	**81.1**	**0.4672**

Bold fonts indicate the best results.

Our Framework Design: To verify the effectiveness of the design choice of our proposed framework, we have conducted an ablation study by comparing our method (i.e., BPNN + KDM) with two alternative combinations, namely, (1) BPNN + decoder and (2) InceptionV3 + KDM.

Essentially, we attempt to replace our encoder—BPNN with the deep learning model InceptionV3 to test whether BPNN is a better choice of encoder to process color themes and relevant information. We also replace our KDM with a decoder of three **full connection (FC)** layers to test whether KDM is a better choice for user preference rating regression. Specifically, the input of the decoder is the rating prediction of a color theme from BPNN, while the output is a user-specific color theme rating. The three FC layers constitute the decoder, composed of 128 nodes, 64 nodes, and 32 nodes, respectively, which can generate the features for the color theme rating task. Following each FC layer, batch normalization is added to improve generalization and accelerate the convergence rate of training. We did not perform any fine-tuning, but the model is re-trained to predict the ratings.

Design Comparison: As shown in Table 1, the design choice of BPNN + kernel density distribution model achieves a better accuracy than the alternative BPNN + decoder scheme (by 31.9% for RMSE, and by 23.8% for MAE). Using BPNN as the encoder to process color themes and relevant information is more accurate than using InceptionV3 (by 4.2% for RMSE, and by 4.9% for MAE). This result implies that, unlike dealing with image processing tasks, deep neural networks, such as InceptionV3, are inferior options to BPNN when dealing with simple five-color theme features.

Since we use majority voting to quantize the continuous values of regression into discrete values, we use categorical quantitative metrics to compare the performances of different methods. In this work, we select Acc (Accuracy) and **Pearson linear correlation coefficient (PLCC)** as the two used categorical quantitative metrics.

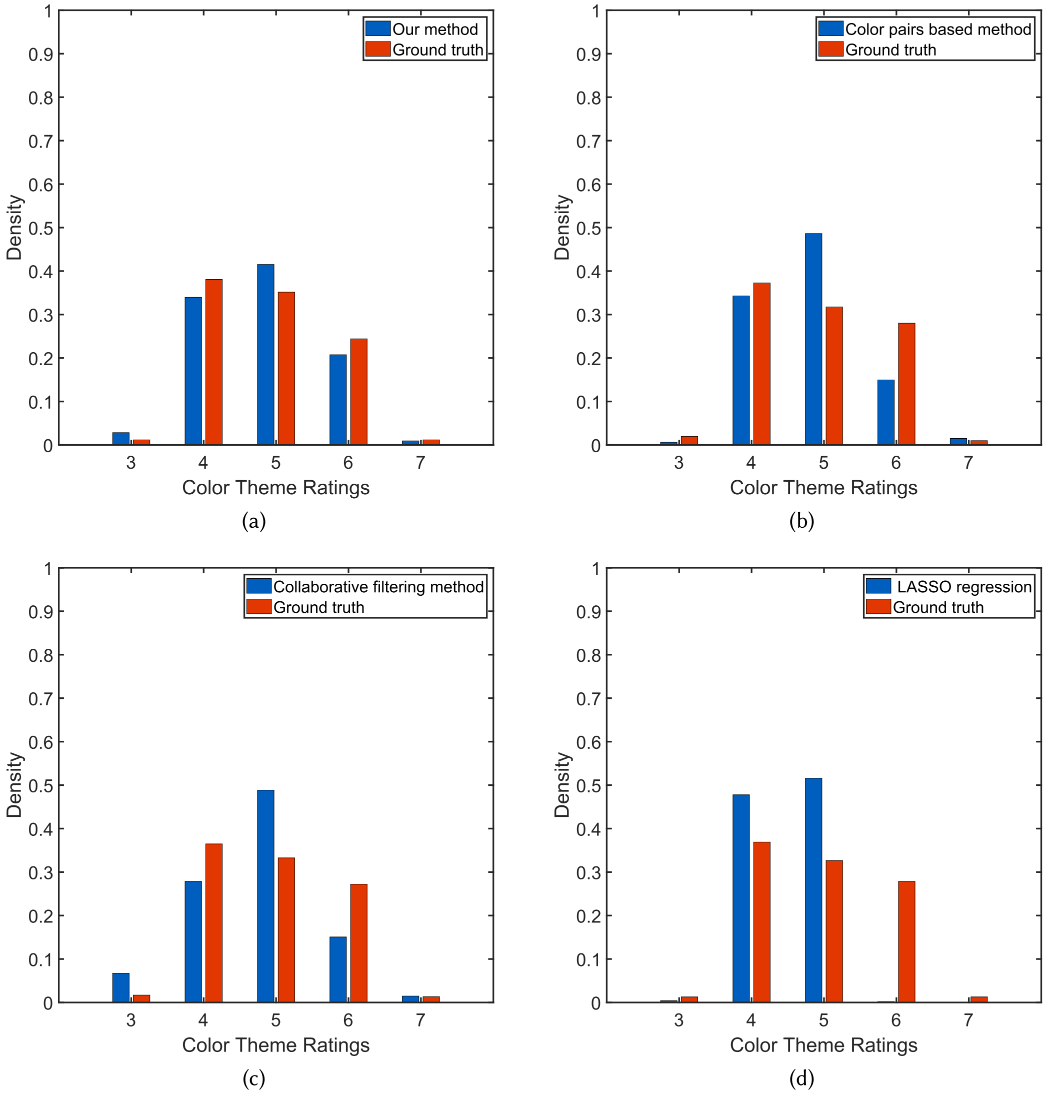
Table 2. Color Theme Evaluation Performances of Different Methods Using Different Regression and Categorical Quantitative Metrics on the MTurk Dataset

Model \ Metric	RMSE ↓	MAE ↓	Acc(%) ↑	PLCC ↑
LASSO regression based [40]	1.2125	0.8697	51.7	0.1175
Color pairs based [68]	0.9852	0.7124	60.1	0.2447
Collaborative filtering [41]	0.8301	0.6322	63.2	0.3151
Alternative combination (BPNN + decoder)	1.0248	0.7412	57.4	0.2257
Alternative combination (InceptionV3 + KDM)	0.7765	0.5921	74.3	0.3506
Our method	**0.7074**	**0.5852**	**77.4**	**0.4255**

Bold fonts indicate the best results.

Acc is one of the widely used metrics for evaluating the performance of classification, and is computed by $Acc = \frac{TP+TN}{P+N}$, where P and N denote the number of positive and negative samples, respectively; TP and TN are the number of correctly classified positive and negative samples. Specifically, since our collected dataset employs a seven-point scale, we consider ratings of no smaller than four points in our collected dataset as positive samples, and ratings of smaller than four points as negative samples. Also, since the MTurk dataset employs a five-point scale, we consider ratings of no smaller than three as positive samples, and those smaller than three as negative samples. PLCC can be used to evaluate the correlation between the predicted ratings and the ground truth after majority voting. As can be seen from Table 1, with the Acc metric, the collaborative filtering based method achieved the highest accuracy, 74.9%, among the existing methods. By contrast, our method achieved 81.1% accuracy, which significantly outperforms all the existing methods. By comparing the Acc (or PLCC) of our method and that of alternative combinations (BPNN + decoder and InceptionV3 + KDM), we can see our design choice of BPNN + KDM significantly boosts the Acc (or PLCC) performance of our proposed model. To further illustrate the generality of our model, Table 2 shows the accuracy performance of our method and state-of-the-art methods on the MTurk dataset [41]. Although the Acc performance of our method is slightly reduced (i.e., from 81.1% on our collected dataset to 77.4% on the MTurk dataset), our method still achieved the best accuracy among all the methods in comparison. Also, we can observe similar comparison results with respect to the PLCC metric, and our method achieved the best PLCC performance among all the methods.

Distribution of Color Theme Ratings: To have an in-depth analysis on the accuracy of our method, we compared the probability distributions of the color theme ratings generated by different methods against the ground truth. The results are shown in Figure 10. We can see that both the Color Pairs based method [68] (Figure 10(b)) and the LASSO regression based method [40] (Figure 10(d)) produce similar distributions, where most of the predicted color theme ratings are concentrated at ratings of 4–5. However, the ratings by the Color Pairs based method [68] also distribute to other intervals, while the LASSO regression based method [40] cannot achieve that. This is because general color theme evaluation models cannot predict the ratings of different users on the same color theme. Therefore, in order to reduce the prediction error during model training, they tend to be over-fitted in the intervals of 4–5. The collaborative filtering based method [41] (Figure 10(c)) worked significantly better than the above two methods [40, 68] by producing a distribution of the predicted color theme ratings, which can better match with that of the ground truth, because this method can infer user variations through matrix factorization. Finally, our method (Figure 10(a)) generated the most accurate distribution among all the four methods, with respect to the probability distribution of the ground truth. This validates that taking into account the variations of user aesthetic preference is a critical factor to achieve accurate color theme rating predictions.



(a)

(b)

(c)

(d)

Fig. 10. Comparison of the predicted and ground-truth color theme rating distributions for different color theme evaluation models, including (a) our method, (b) the color pairs based method [68], (c) the collaborative filtering based method [41], and (d) the LASSO regression based method [40].

We used Kullback-Leibler Divergence and squared earth movers distance to calculate the difference of the probability distributions between the ground truth and the color theme ratings predicted by different methods in Figure 10. Additionally, we analyze the correlations between different probability distributions using Pearson (ρ_P) and Spearman correlation coefficients (ρ_{SP}), with the results presented in Table 3. The rating distribution obtained by our method has significantly smaller errors than those by other methods, compared to the ground-truth rating distribution.

Extreme Test Data Challenge: To thoroughly validate the robustness of our method, we challenged different methods by using some test cases with extreme color preference ratings. Since most of the color theme ratings in

Table 3. Quantitative Comparison of Evaluation Models for Different Color Themes in Figure 10 Data

Model \ Metric	KL	EMD	ρ_P	ρ_{SP}
Our method	**0.0127**	**0.0068**	**0.8887**	**0.8903**
Color pairs based [68]	0.5407	0.5186	0.4390	0.5905
Collaborative filtering [41]	0.1055	0.3079	0.7675	0.8532
LASSO regression based [40]	1.4195	0.7889	0.1129	0.3536

Bold fonts indicate the best results.

Table 4. RMSE and MAE Comparison with Extreme Test Data

Model \ Metric	RMSE	MAE
Our method	**0.6536**	**0.5844**
Color pairs based [68]	0.8725	0.8271
Collaborative filtering [41]	1.3523	1.2425
LASSO regression based [40]	4.3345	2.0642

Bold fonts indicate the best results.

our dataset are in the range of 4–6, we randomly selected 500 sparse color theme ratings in the ranges of 1–3 and 6–7 as an additional test set to perform comparisons. We again compared all the methods through the RMSE and MAE metrics, and the newly obtained results are shown in Table 4. Notably, our method outperformed all the state-of-the-art methods and achieved a comparable accuracy even with extreme color theme preference ratings as the test data, compared to the accuracy of our method as reported in main text Table 3. Meanwhile, it can be observed that the LASSO regression based method [40] cannot properly handle such a challenge and had a big drop in accuracy. More importantly, although the collaborative filtering based method [41] can infer user variations through matrix factorization, its performance significantly deteriorated under this challenge. This suggests that the collaborative filtering based method [41] may only work well if the color theme ratings could approximately follow a normal distribution. By contrast, our method explicitly models individual users' aesthetic preferences. Therefore, it offers a more robust solution to color theme evaluation even for extreme test data.

Robustness to Personalized Datasets: Being able to infer user aesthetic preference is a unique feature of our method. Hence, we performed a test to see whether our method can function properly if all color themes were rated by users with the same aesthetic preference. This serves as another test to validate the robustness of our method. To carry out the test, we constructed a personalized dataset based on all the color theme ratings done by the user #34667 in our original dataset. We chose this user because the user evaluated 2,233 color themes, producing a total of 7,138 color theme ratings, and the user rated some color themes multiple times. This gave us a sufficient amount of data to train our model. Figure 11 visualizes our findings by showing how each color theme is rated by (1) our method with personalized dataset adaption, (2) the user #34667 (obtained by majority voting if multiple ratings exist), and (3) our native method (i.e., trained by using the dataset as described in main text Section 4). As shown in the figure, the predicted color theme ratings generated by our method with personalized dataset adaption are generally in line with the ratings given by the user #34667, while the predicted ratings by our native method could be quite different from those given by the user #34667. This demonstrates that our method can properly predict color theme ratings even if the training dataset was significantly biased toward a specific user.

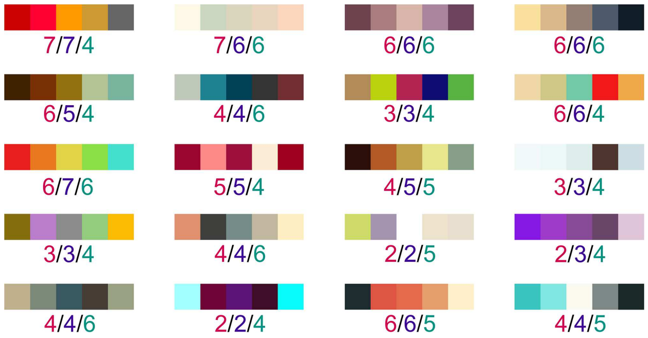


Fig. 11. Comparison of the color theme evaluations. The three numbers (from left to right) underneath each color theme are the color theme ratings given by: (1) our method with personalized dataset adaption (red), (2) the user #34667 (purple), and (3) our native method (dark cyan), respectively.

6.2 Personalized Color Theme Assessment

To quantify the performance of our personalized aesthetics model, we evaluated our approach using a baseline method as well as compared it with the collaborative filtering based approach [41].

We first leveraged a training set of our dataset to learn a multi-user personalized evaluation model, and then conducted personalized aesthetic experiments on a test set. Specifically, we screened our dataset to ensure that the number of color themes labeled by each annotator is more than 105 to facilitate subsequent experiments. Based on this criterion, we selected 22,851 color themes labeled by 396 annotators (called “training users”) as a training set, and selected 4,016 color themes labeled by 57 annotators (called ”test users”) as a test set. Further, we ensure the in-existence of overlaps between the training users and the test users. This allows us to simulate situations where each user only provides ratings on his or her own color themes, and algorithms cannot access those color themes and ratings in the test set beforehand. In a multi-user personalized evaluation model, the ranking consistency between the predicted and ground-truth results is an important evaluation criterion [24, 32, 48, 71]. We employ the s**pearman rank-order correlation coefficient (SROCC)** [38] to evaluate the performance of multi-user personalized evaluation approaches.

Baseline Method: For each test user, from the test set we randomly select k color themes that are labeled by this specific test user, and further add them into the training set. Meanwhile, we use the remaining color themes rated by the same test user (i.e., excluding the k selected color themes) in the original test set as a new test set. Then, we use the expanded training set to train our approach and perform testing on the new test set. Due to the randomness of selecting the k color themes in the above process, we ran the experiments 50 times for each test user, and reported the averaged results as well as the standard deviation. In order to evaluate the performance of our approach for different scenarios, we chose $k = 10$ and $k = 100$, respectively. Figure 12 shows the comparison between the ground-truth personalized ratings of three randomly selected test users and the prediction results by our approach when $k = 100$. As can be seen from the figure, the ratings predicted by our personalized color theme evaluation model are very close to the ground-truth, user-specific ratings on the color themes. This shows that our model can robustly learn the aesthetic preferences of specific users, with the aid of a small number of training samples rated by the specific users.

Comparison with Existing Methods: Since the collaborative filtering based method [41] also learns the users' subjective preferences for color theme evaluations, we compared our personalized aesthetics model with the collaborative filtering based method using the same baseline process (as described above). For the sake of completeness, we also added two existing general (i.e., without personalization) color theme preference prediction models [40, 68] and two ablation studies (BPNN + decoder and InceptionV3 + KDM) to this comparison.

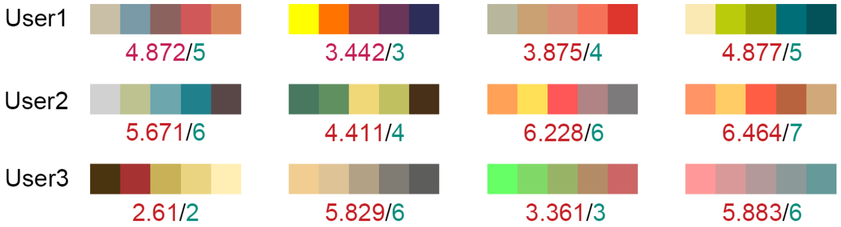


Fig. 12. Comparison between the ground-truth personalized ratings of three randomly selected test users and the prediction results by our approach when $k = 100$. The predicted personalized color theme rating by our method (red) and the ground-truth rating by the user (dark cyan) are shown below each color theme.

Table 5. SROCC Comparison of Results between Our Method and the State-of-the-Art Methods on Our Dataset

Method	$k = 10$	$k = 100$
LASSO regression based [40]	0.1532 ± 0.070	0.1779 ± 0.082
Color pairs based [68]	0.2871 ± 0.042	0.2944 ± 0.049
Collaborative filtering [41]	0.3477 ± 0.004	0.3615 ± 0.008
Alternative combination (BPNN + decoder)	0.1746 ± 0.066	0.2015 ± 0.072
Alternative combination (InceptionV3 + KDM)	0.3501 ± 0.006	0.3822 ± 0.010
Our method	**0.3954 ± 0.004**	**0.4257 ± 0.012**

Bold fonts indicate the best results.

Table 5 shows the comparison results on our dataset in terms of SROCC. We can see that the general color theme evaluation methods [40, 68] do not work well for this case (i.e., SROCC values are low), because they do not consider the aesthetic differences of distinct users. The collaborative filtering based method [41] outperformed the above two general models by predicting user preferences, but its performance improvement is limited. When k is increased from 10 to 100, the collaborative filtering based method only has a marginal improvement (from 34.77% to 36.15%, that is, 1.38% improvement). By contrast, as shown in Table 5, our method has a larger SROCC improvement (i.e., 3.03% improvement) than the collaborative filtering based method when k is changed from 10 to 100. This verifies the effectiveness of our kernel probability model for user aesthetic preference modeling. In the ablation study, we still experimented three FC layers as the decoder instead of the KDM, and the results were only slightly better than the collaborative filtering based method [41], indicating that the decoder module does not have the ability to optimize the personalized task. We also experimented the deep learning model InceptionV3 instead of BPNN for the encoding task, our model's SROCC is better than InceptionV3 + KDM.

To further look into the effectiveness of our method for learning individual-specific aesthetic preferences, we first directly used our BPNN model to predict the averaged ratings on color themes and then further computed the SROCC prediction performance of the 57 test users in our test set. After that, we calculated the SROCC increase by using our personalized color theme evaluation model with $k = 100$, as described in the above baseline process. As shown in Figure 13, we observe that the SROCC performances for almost all the 57 test users increase significantly through our personalized model. For example, the SROCC performance of the user #41 improves from 0.22 to 0.60 (that is, an impressive improvement of 0.38). The average SROCC improvement of all the 57 test users is 0.19 (the range is from 0.235 to 0.426). Therefore, our personalized color theme evaluation model can effectively make use of individuals' aesthetic preference to predict specific users' color theme evaluations through the introduced kernel density estimation model.

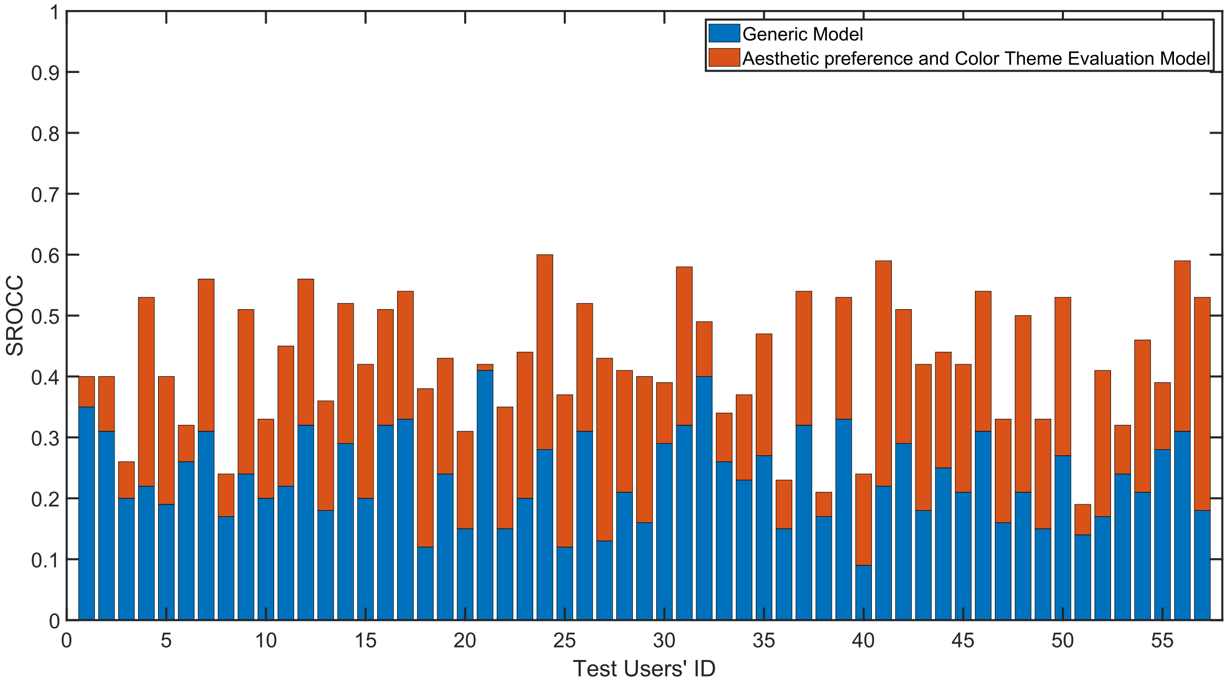


Fig. 13. SROCC performance improvements of the 57 test users by directly comparing the generic color theme evaluation model (i.e., our BPNN model) and our personalized aesthetic color theme evaluation model when $k = 100$. The blue bars show the SROCC values of the 57 test users by our generic BPNN model, and the red bars show the SROCC increase after our personalized color theme evaluation model is used.

Performance Comparison on the MTurk Dataset: In order to further validate the effectiveness of our personalized model on the MTurk dataset, we need to first filter the MTurk dataset to ensure no duplicated color themes between the test set and the training set. To this end, we selected 193 annotators who rated a total of 7,978 color themes. The number of color themes labeled by each annotator is between 106 and 200 (the average value is 151). We selected 5,915 color themes labeled by 159 annotators as the training set, and 2,063 color themes labeled by 34 annotators as the test set.

Only the collaborative filtering based method [41] publicly released its experimental results on the MTurk dataset. To validate the performance of our personalized method on the MTurk datasets, we compared our method with the collaborative filtering based method. We used the color themes in the training set to train both methods, following the above baseline process. In this experiment, we set $k = 10$ and $k = 100$, respectively. The average results and the standard deviations of 50 repeated experiments are reported in Table 6. As can be seen from this table, the performance of our method is clearly better than that of the collaborative filtering based method on the MTurk dataset for both $k = 10$ and $k = 100$.

Cross-dataset Evaluation: A model can work well in practical applications if the model could be generalized to unseen data, i.e., high in generalization ability. To demonstrate such an ability of our proposed model in performing personalized evaluation tasks, we have trained our model with one dataset and test it with another dataset. We have also performed the same experiments with two other existing methods, including Collaborative Filtering [41] and Color Pairs based method [68], for comparison.

Specifically, we conduct a cross-dataset evaluation. To compare the generalization abilities of different methods, in one set of experiments, we train each method with our proposed dataset (as known users for personalized

Table 6. SROCC Comparison of Results of Our Method and the Collaborative Filtering Based Method [41] on the MTurk Dataset

Method	k = 10	k = 100
Collaborative filtering [41]	0.3537 ± 0.005	0.4782 ± 0.016
Our method	**0.4112 ± 0.006**	**0.5155 ± 0.010**

Bold fonts indicate the best results.

Table 7. Cross-Dataset Evaluations on Different Methods

Method		Our method		Collaborative filtering		Color pairs	
Holistic analysis		**stiff: 0.4045**		stiff: 0.3545		stiff: 0.2809	
		stable: 91.53%		stable: 88.76%		stable: 85.73%	
SROCC	Train Test	Ours	MTurk	Ours	MTurk	Ours	MTurk
	Ours	0.4159	**0.3801**	0.3559	0.3049	0.2902	0.2388
	MTurk	**0.3535**	0.4686	0.3094	0.4476	0.2265	0.3681

Both holistic analysis based on stiffness and stability and fine-grained results based on SROCC are reported. Bold fonts indicate the best results.

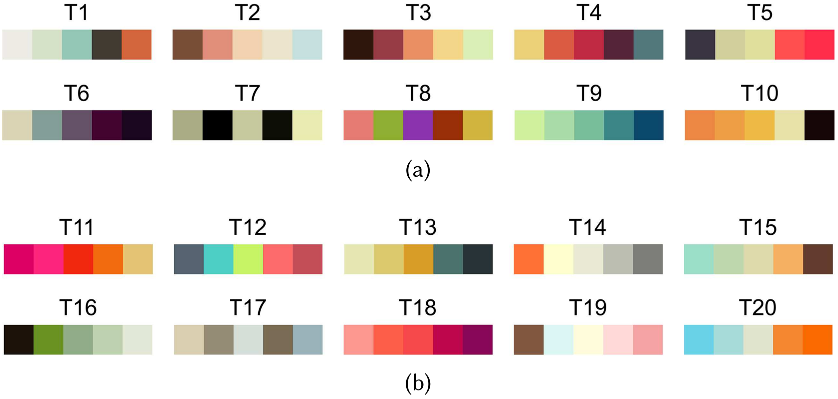
evaluation tasks) and test with MTurk (as unknown users), and vice versa, i.e., train with MTurk and test with our proposed dataset in another set of experiments. The sizes of the training and test sets are the same as those in the previous experiments. Since the two datasets have different rating ranges for the ground truth, we have performed a normalization operation on the ground truth.

Table 7 shows both holistic analysis and fine-grained results of our cross-dataset evaluation on all methods cross-dataset evaluation results of different methods. For holistic analysis, we have adopted two cross-dataset measurements by [11], namely stiffness and stability, to evaluate the overall performance of different methods on cross-datasets. Stiffness reflects the absolute performance of a system under cross-dataset setting. Given a methods, the result of its cross-dataset evaluation using two datasets yields a 2 × 2 matrix with its elements denoted by U_{ij}, its stiffness is therefore $r^{\mu} = \frac{1}{N*N} \sum_{i,j} U_{ij}$. On the other hand, stableness characterizes the relative performance gap between in-dataset and cross-dataset test. It is calculated as $r^{\sigma} = \frac{1}{N*N} \sum_{i,j} (U_{ij}/U_{jj} \times 100)$.

As shown in Table 7, both the stiffness and stableness of our method are higher than those of the other two methods. In addition, the fine-grained results based on SROCC also show that our method outperforms the other two methods. In conclusion, our methods is more stable across datasets and has better generalization ability.

6.3 User Study

We conducted a user study to evaluate the effectiveness of our personalized color theme evaluation method based on 20 randomly selected color themes from the COLOURLovers dataset, as shown in Figure 14. The selected color themes cover different hues, saturations, and brightnesses. We invited 100 participants (50 males and 50 females, with the ages between 20 and 65) to evaluate these color themes. These participants lacked professional expertise in visual perception or color science. Instructions were given to the participants, informing them to use their color aesthetics. They were also asked to select a rating from 1 to 7 scale to indicate their preference level on color themes. To test our personalized color theme evaluation model, we used the ratings on the ten color themes in Figure 14(a) by each participant as additional training data, together with the default training data as describe in Section 4, to train a personalized color theme evaluation model for this specific participant. Figure 14(b) show



(a)

(b)

Fig. 14. The 20 selected color themes in our user study. (a) 10 color themes are used as additional data for training personalized models, and (b) the other 10 color themes are used for testing.

the remaining ten color themes for testing our personalized models in the user study. This study was certified by the university committee for use with human participants, and all participants understood and harmonized the content of the informed consent form.

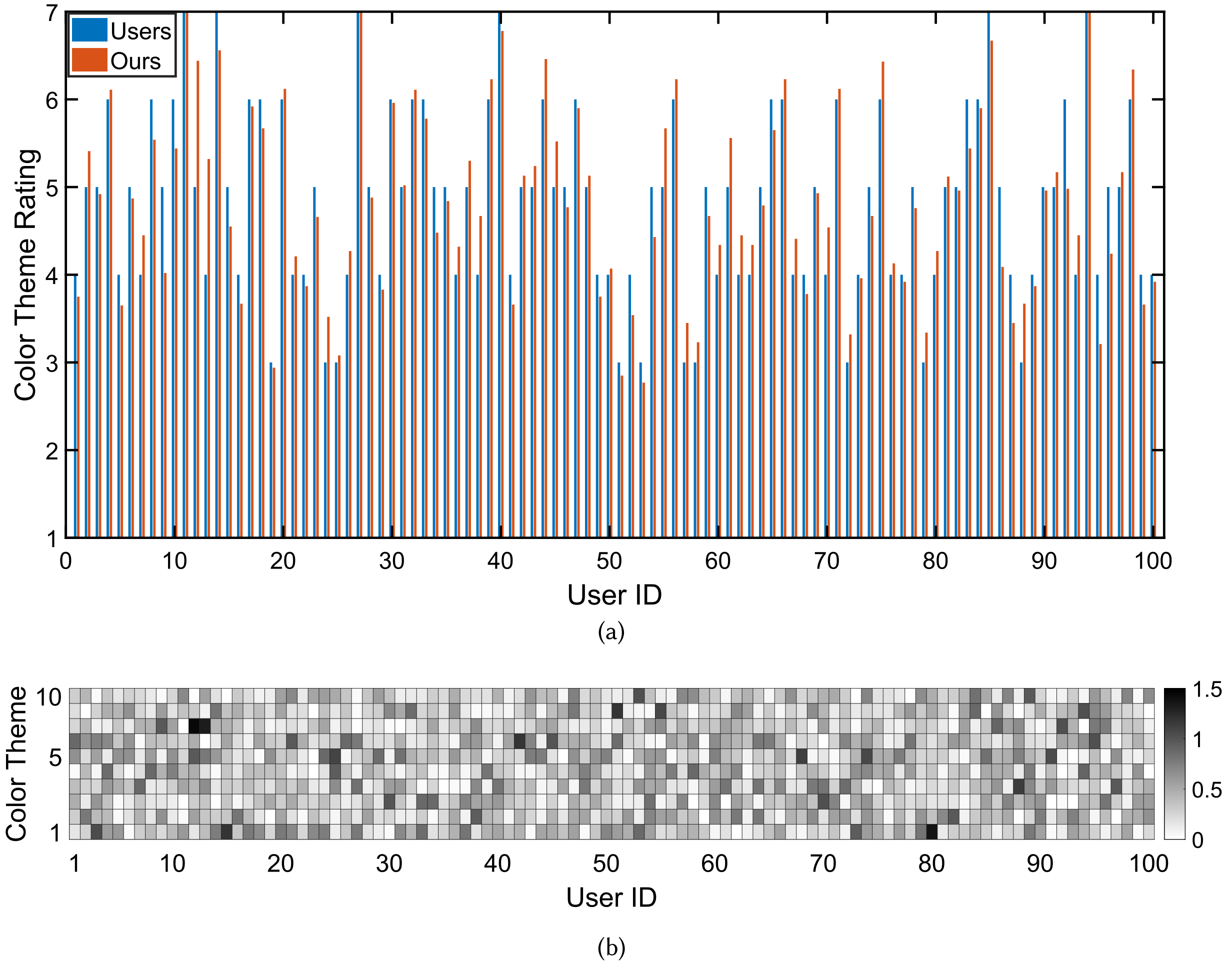
Figure 15(a) shows the ratings of a randomly selected color theme example (i.e., T18 in Figure 14(b)) in our study: the first type of ratings were predicted by our personalized method (indicated as "Ours" in the figure), and the second type of ratings were rated by all the participants based on their own preferences (indicated as "Users" in the figure). Since our method learned a different personalized model for each participant, in Figure 15(a) we can see the predicted ratings even for the same color theme (T18 in this case) are different by our personalized models that were customized for different participants. However, as shown in the figure, the ratings predicted by our personalized method are sufficiently close to those made by the participants. This is further evidenced by the evaluation results of our personalized method's predicted scores under different quantitative metrics, as shown in Table 8.

Figure 15(b) shows the heat map visualization of the differences between the predicted ratings by our personalized method and the ratings given by the participants for all the 10 test color themes. In this figure, the X-axis represents user IDs and the numbers 1–10 on the Y-axis represents the test color themes T11–T20 in Figure 14(b), respectively. As shown in the figure, most of the color blocks are white or light gray, meaning that, with a large number of data points (100 participants × 10 color themes = 1,000), the predicted ratings by our personalized method are reasonably close to the ground-truth evaluations given by the participants.

Finally, we also compared the color theme ratings generated by our personalized method with the ground-truth color theme ratings given by participants, using a group-based comparison theme. We first used our model to quantify the aesthetic preferences of the participants and divide the range of the user aesthetic preference $\{\alpha_i\}$ into three groups: [0, 0.33), [0.33, 0.67), and [0.67, 1.0]. Then, for the participants grouped into each group, we conduct a quantitative comparison between the scores generated by our approach and those provided by participants. Table 9 presents the comparative results for various quantitative metrics across different groups. Here, ρ_P represents Pearson correlation, and ρ_{SP} represents Spearman correlation. The results indicate that regardless of whether participants tend to underestimate or overestimate the aesthetic quality of color themes, the color theme ratings predicted by our method are highly correlated with the actual ratings of the majority of participants, exhibiting no bias toward any specific group of participants based on aesthetic preferences.

6.4 Correlation between Aesthetic Preference and Harmony

In this article *user aesthetic preference* refers to the difference of the user's perception on color aesthetic, which is a subjective judgment on how the user perceives colors and color combinations. Meanwhile, *color harmony* is a



(a)

(b)

Fig. 15. (a) The predicted ratings by our personalized method and the ground-truth ratings given by the participants on a specific color theme (T18 in Figure 14). (b) Heat map visualization of the differences between ratings predicted by our personalized method and given by participants for all 10 test color themes.

Table 8. Quantitative Comparison of Predicted Ratings by Our Personalized Method and Personalization Ratings Given by Participants

Metric Method	MAE	RMSE	ρ_P	ρ_{SP}
Our method	0.5526	0.6254	0.8144	0.8303

mixed bag of both objective and subjective factors measuring how colors get along with each other. Hence, these are inter-related concepts. Some previous studies investigated the relationship between user preference and color harmony qualitatively. Schloss and Palmer [54] conducted a comprehensive study to understand this relationship by performing experiments against color pairs. A major finding of their study is that there is a high correlation

Table 9. Quantitative Comparison of Ratings by Groups in Our User Study

Metric Group	MAE	RMSE	ρ_P	ρ_{SP}
[0,0.33)	0.6032	0.6923	0.8126	0.8630
[0.33,0.67)	0.5013	0.6318	0.8532	0.8379
[0.67,1]	0.4781	0.5934	0.8375	0.8696

Participants are divided into three groups according to their user aesthetic preference values: (a) $\{\alpha_i\} \in [0, 0.33)$, (b) $\{\alpha_i\} \in [0.33, 0.67)$, (c) $\{\alpha_i\} \in [0.67, 1]$.

between user preferences and the color harmonies of color pairs (r = +0.79), where r is the Pearson correlation coefficient.

Inspired by the above problem, we study whether we can obtain similar or different findings on the correlation between the user aesthetic preferences and the color harmonies of five-color themes from the perspectives of both the average ratings and personalized ratings. We used some color theme data in O'Donovan et al. [40], which were also collected from the COLOURLovers website, and selected the same color themes in our dataset. The subjective ratings of the selected color themes, obtained via large-scale MTurk in O'Donovan et al. [40], were regarded as the ground truth color harmony. To this end, we screened a total of 8,563 color themes (included in both the dataset in O'Donovan et al. [40] and our dataset in this work) evaluated by 863 annotators, and further used majority voting to integrate the ratings on the same color themes by different annotators, so that the correspondences between the user ratings and the color themes are one-to-one. We normalized the color harmony ratings so that their range is from 1 to 7.

Figure 16(a) shows the scatter plot of the ground-truth color harmony ratings from O'Donovan et al. [40] (X-axis) and the predicted color theme ratings through majority voting by our method (Y-axis). The red dotted line in the figure is the regression line fitted to the scatter plot, and its slope is 0.787. The calculated Pearson correlation coefficient r is +0.81. The result validates that there is a high positive correlation between the two types of ratings, although there are some differences. For example, we found about 16% of the color themes are underestimated, and 4% of the color themes are overestimated. By further analyzing the ground-truth user ratings on these under-estimated/over-estimated color themes, we found that the number of annotators who rated these color themes is usually fewer than 10. Meanwhile, these ratings are not normally distributed and cannot be extracted well by majority voting.

In order to further investigate the relationship between the personalized ratings by our method and the ground-truth color harmony ratings in O'Donovan et al. [40] for users with different aesthetic preferences, we looked into annotators with different aesthetic preferences (α values) and their color theme evaluations. As shown in Figure 16(b), we set α = 0.8, 0.5, and 0.2, respectively, to represent high, medium, and low levels of aesthetic preferences. As shown in the figure, under the same test conditions, users with high (or low) aesthetic preference usually give higher (or lower) ratings on color themes than the ground-truth ratings. When the aesthetic preference of users is an appropriate value (i.e., α = 0.5), the color theme ratings given by users are highly consistent with the ground-truth ratings. These findings are consistent with those in previous studies [54], but our findings are based on more generalized five-colors themes rather than simple color pairs in Schloss and Palmer [54]. In theory, when α = 0.5, the personalized ratings should approximately equal to the ground-truth color theme ratings, but experimental results always produce certain deviations (r = +0.81) due to the unavoidable human biases. We also computed RMSE and MAE metrics as in Table 10 for the three cases in Figure 16(b). These numbers show that as the aesthetic preference α is close to 0.5, the error between the predicted ratings by our methods and the

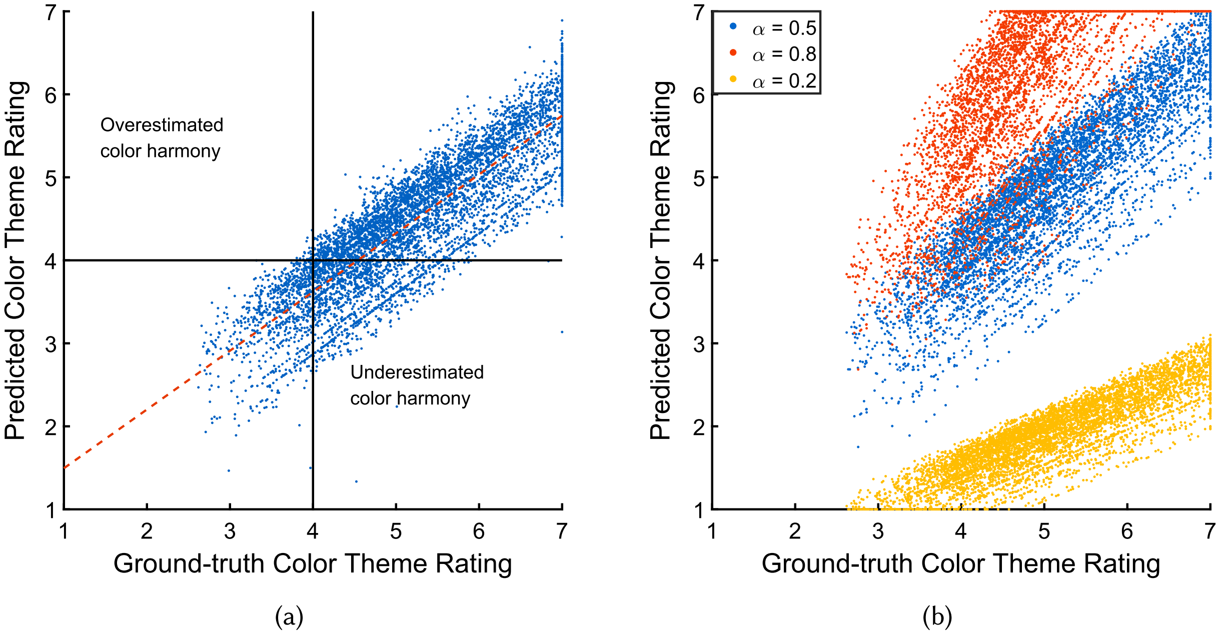


Fig. 16. Correlation plotting of two types of color theme ratings. (a) The plot of the predicted color theme ratings by our method (Y-axis) against the ground-truth color harmony ratings (X-axis) in O'Donovan et al. [40]. The red dashed line shows the fitted regression line ($y = 0.787 + 0.7079x$). (b) Plots showing the influence of different user aesthetic preferences (α values) on the relationship between the personalized ratings by our method (Y-axis) and the ground-truth color harmony ratings (X-axis).

ground-truth ratings becomes smaller. This also suggests that it is generally difficult for users with too high and too low aesthetic preferences to find highly harmonious color themes as their preferred color themes.

We also explored the relationship between user aesthetic preference and color harmony by changing the harmony of five-color themes. As shown in Figure 17, we used a color suggestion method [68] to change both one color of the five-color color themes and the order of the five colors to generate new color themes that are more harmonious or inharmonious than the original color themes. We selected three users from our dataset and used our model to predict their aesthetic preference and personalized ratings for the original and new color themes, and compared the results with the color pairs based prediction model by [68]. From the figure, we can see that when the user aesthetic preference α is approximately 0.5, the user's personalized ratings are also very close to the predicted harmony scores by [68]. Otherwise, the personalized ratings by our method are significantly higher or lower than the predicted harmony scores by [68] when the α increases or decreases. Besides coming up with a more generalized finding on how user preferences (or user aesthetic preferences) and color harmonies are related to five-color themes, we anticipate our results would give fresh insights for designers or relevant professionals to work on more suitable color choices for the artifacts they produce, based on the characteristics of target audiences.

6.5 Selected Applications

In recent years numerous commercial applications have been developed to automatically recommend visual content, such as video (e.g., YouTube^{TM}) and images (e.g., Pinterest^{TM}), to users based on their individual preferences. Taking the Pinterest application as an example, a user first needs to select a few image categories as

Table 10. RMSE and MAE Measurements for User Aesthetic Preference Groups

Metric Aesthetic preference	RMSE	MAE
α = 0.5	0.7388	0.6647
α = 0.8	11.2041	2.8744
α = 0.2	12.4025	3.5709

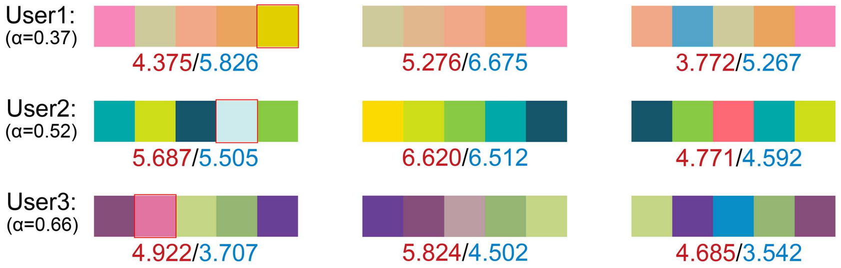


Fig. 17. The comparison of the personalized ratings by our method (red) and the predicted color harmony scores by Yang et al. [68] (blue) for three users with different aesthetic preferences. The left column shows the original color themes in our dataset, and the middle and right columns show the more harmonious and more inharmonious color themes, respectively, obtained by the color suggestion method in Yang et al. [68].

his/her categories of interest, and also pick a few images from the categories of interest as his/her favorite images. The Pinterest App can then learn the user preference and later recommend and push new images to the user based on his/her individual-specific preference. Since our color theme evaluation model is based on the modeling of individual users' aesthetic preferences, our model can be potentially applied for such individual-customized recommendation applications. We will illustrate the utility of our approach and dataset in the field of personalized recommendations, specifically in the contexts of general images, fashion, and interior design.

6.5.1 Personalized Image Recommendation. To validate the effectiveness of our model for the general image recommendation applications, we conducted the following experiment. We selected images from the AVA database [37] that contains more than 250,000 labeled aesthetic images and 66 semantic categories. We screened out 60 semantic categories from the AVA database to ensure that each image category contains more than 100 images. We invited 30 participants (15 males and 15 females, ages 20–30) to participate in this experiment. These participants did not possess professional visual perception or color science experience.

To narrow down the scope of the experiment, we asked each participant to first select his/her favorite image category from the 60 semantic categories. We then randomly selected 27 images from this category and show them to the participant. Meanwhile, the participant rated these images (refer to Figure 18) based on their aesthetic preferences. In the rating process, the participant was instructed to put more emphasis on the color harmony of the images. For each of the selected images, we also used the algorithm in Lin and Hanrahan [26] to extract its five-color theme.

Based on the extracted color themes and the obtained individual-specific ratings on the 27 pages, we used our approach to model each participant's aesthetic preference value and then further predicted individual-specific ratings on the remaining images in the chosen image category of interest. After that, we selected the top 30 images based on the predicted individual-specific ratings and presented the 30 images to the same participant. Figure 19(a) shows the presented image examples for three participants in our experiment. The participant can choose to "like" a presented image by simply clicking it if he/she indeed likes it, or otherwise just ignore it if

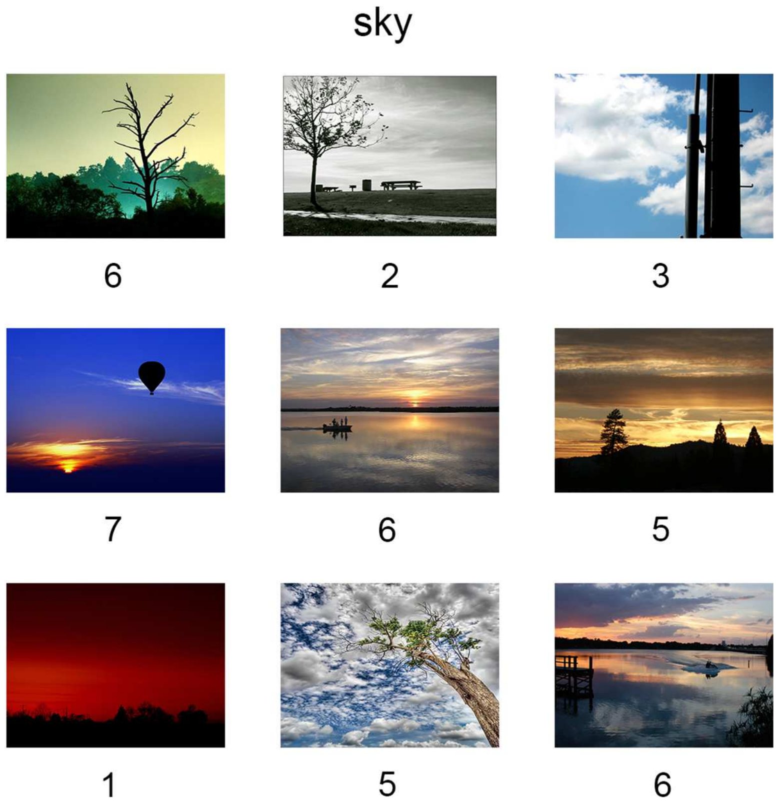


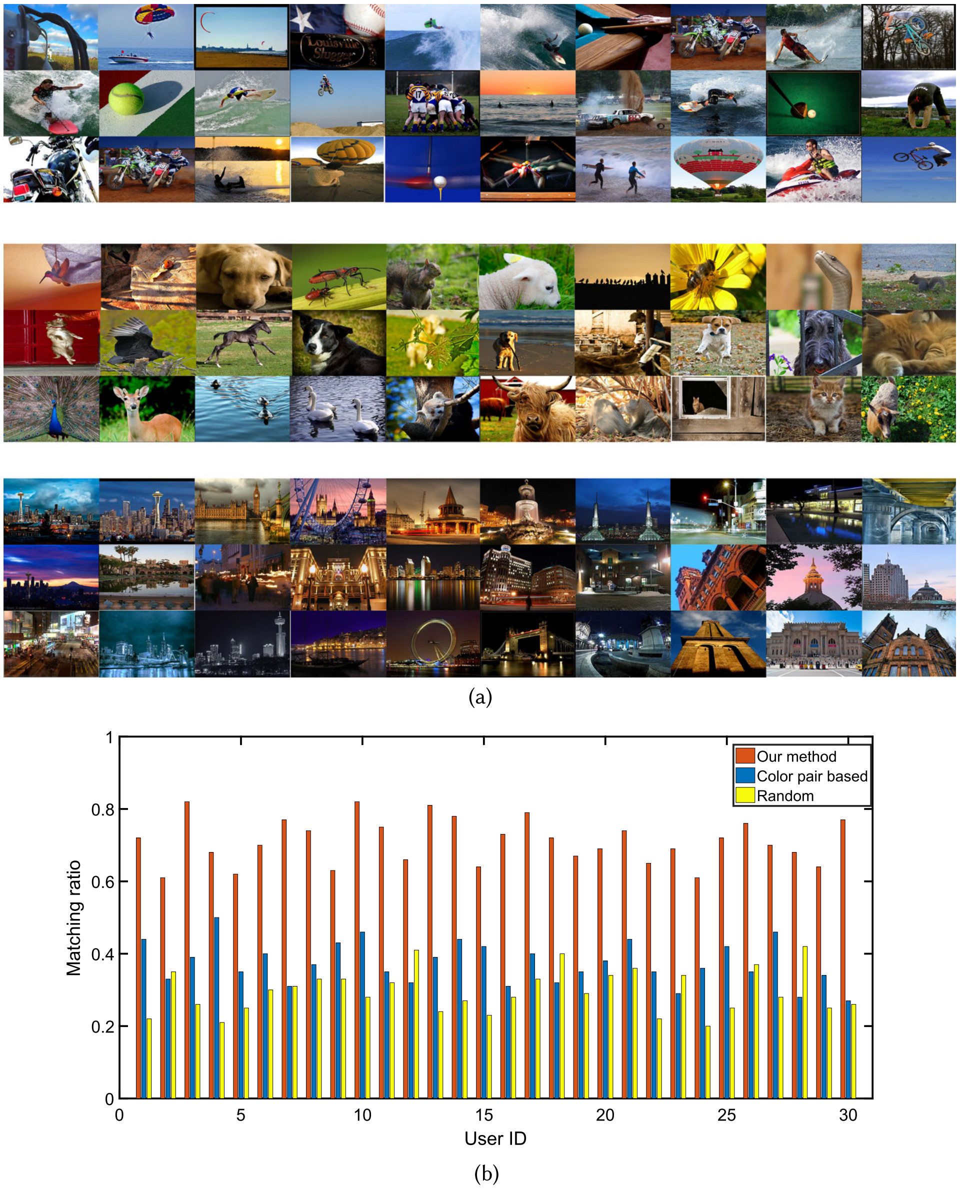
Fig. 18. Some rated image examples in the "sky" semantic category of the AVA database. The score below each image is rated by the participant #7.

he/she is less enthusiastic for it. To this end, we can compute a ratio of the particular participant (called the "ratio of likes" in this writing) as the number of "likes" clicked by the participants divided by 30 (i.e., the number of the presented top-30 images), and the ratio of likes is between 0 and 1.

In this experiment, we also compared our method with the color-pair based method [68] and a random selection method (as the baseline) for doing the same task as described above. Specifically, in the random selection method, we just randomly selected 30 images from the remaining images (i.e., excluding the 27 images manually rated by the specific participant) in the participant's favorite category and presented them to the participant. Figure 19(b) shows the plotting of the ratios of likes for all the 30 participants by the three different methods (i.e., our model, the color-pair based method [68], and the random selection method). As shown in this figure, the color-pair based method clearly achieves higher ratios of likes than the random selection method for most of the participants. Also, our method achieves significantly higher ratios of likes than the other two methods for all the 30 participants.

We are aware that many possible factors can influence such individual preference based recommendation applications. However, we believe color harmony could be one useful factor to consider in order to improve individualized preference based visual content recommendation applications.

6.5.2 Personalized Fashion Outfit Recommendation. We employed the Polyvore Dataset [17, 72] for personalized fashion outfit, consisting of 127,219 fashionable items segmented from 21,889 outfit images, spanning a total of 20 semantic categories. To adapt it for our model, we grouped these 20 fashion items into five categories: outwears, tops, bottoms, shoes, and accessories. We specified that each outfit composition consists of these five fashion



(a)

(b)

Fig. 19. (a) Examples of the top-30 presented images by our method for three participants. The images (from top to bottom, from left to right) are sorted by the predicted color harmony rating in a descending order. The image categories from top to bottom are sports, animals, and cityscape. (b) The plotting of the ratios of likes by the three different methods (our method, the color pair based method [68], and the random selection method) for all the 30 participants in our experiment.

item categories and utilized the algorithm outlined in Lin and Hanrahan [26] to extract the main color of each item, constructing their respective five-color themes.

We invited 30 participants (15 males and 15 females, aged 20–30 years) to take part in the experiment. These participants lacked professional visual perception or color science experience. To streamline the experiment,

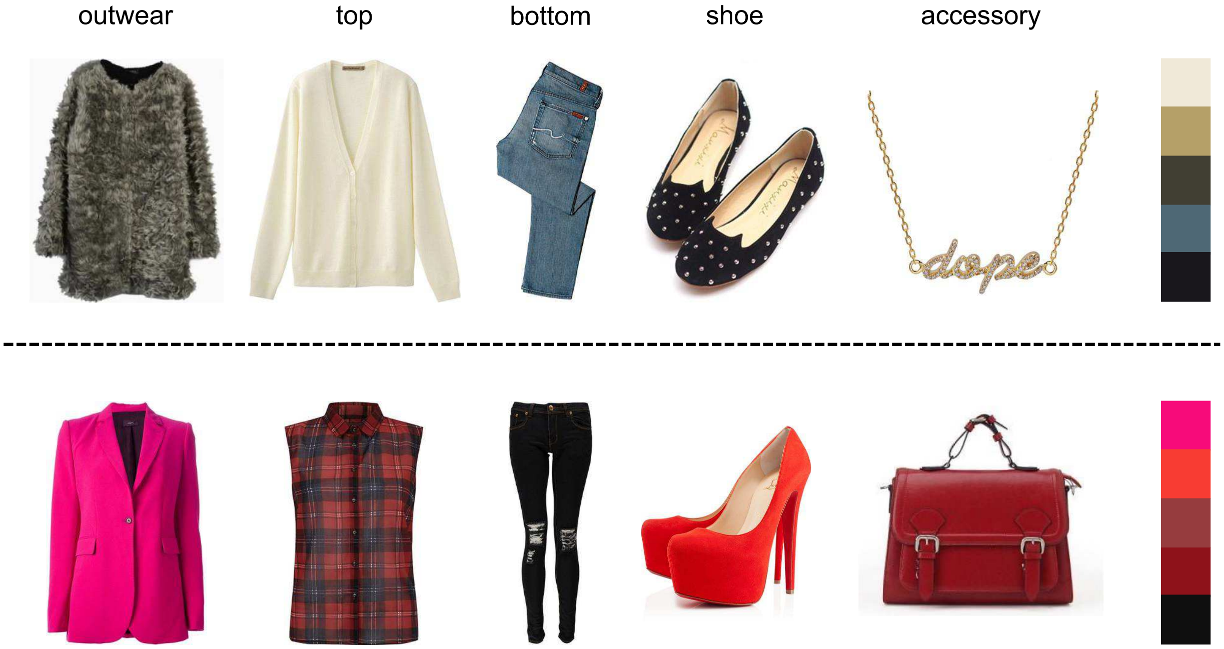


Fig. 20. Left: Examples of the fashion outfit recommendation based on user color preferences. Right: The color themes composed of the dominant colors of five fashion items. The color sequence has been modified using the method in Yang et al. [68].

for each participant, we initially required them to select their top five favorite categories from the 20 fashion item categories, ensuring that these five categories fall within outwears, tops, bottoms, shoes, and accessories. Subsequently, we randomly selected 27 outfit images from each chosen category, tailored to the participant's gender, and presented them to the participants. Participants rated these images based on their aesthetic preferences. Afterwards, leveraging the extracted color themes and personalized ratings, we modeled the aesthetic preference for each participant using our method, and predicting individualized ratings for the remaining images in the selected category of interest. Finally, we regarded the highest-rated combination as the user's favorite fashion outfit. Figure 20 illustrates instances of our recommended fashion outfit. The figure includes the example of color pairings of outfits preferred by two female users. To mitigate the impact of color harmony on user color preferences, we employed the color suggestion method [68] to alter the sequence of color themes in fashion outfit, maximizing color harmony. It is important to note that our method does not account for user preferences in clothing styles and can only represent a color-reference scheme containing user color preferences. Table 11 shows the comparison results of the color ratings predicted by our method and state-of-the-art methods with the personalized color ratings given by the user. It can be concluded from different quantitative indicators that our method achieves the best performance among all methods.

6.5.3 Interior Design Color Recommendation. We employed the SynthHomes dataset [64] to showcase the applicability of our method in interior design. This synthetic dataset comprises 1,000 RGB images, featuring two types of homes with living rooms, kitchens, bathrooms, and bedrooms. We categorized the images based on the functional attributes of the rooms, selecting 785 images for training and 215 images for testing. Each image was processed to extract five-color themes using the method in Lin and Hanrahan [26].

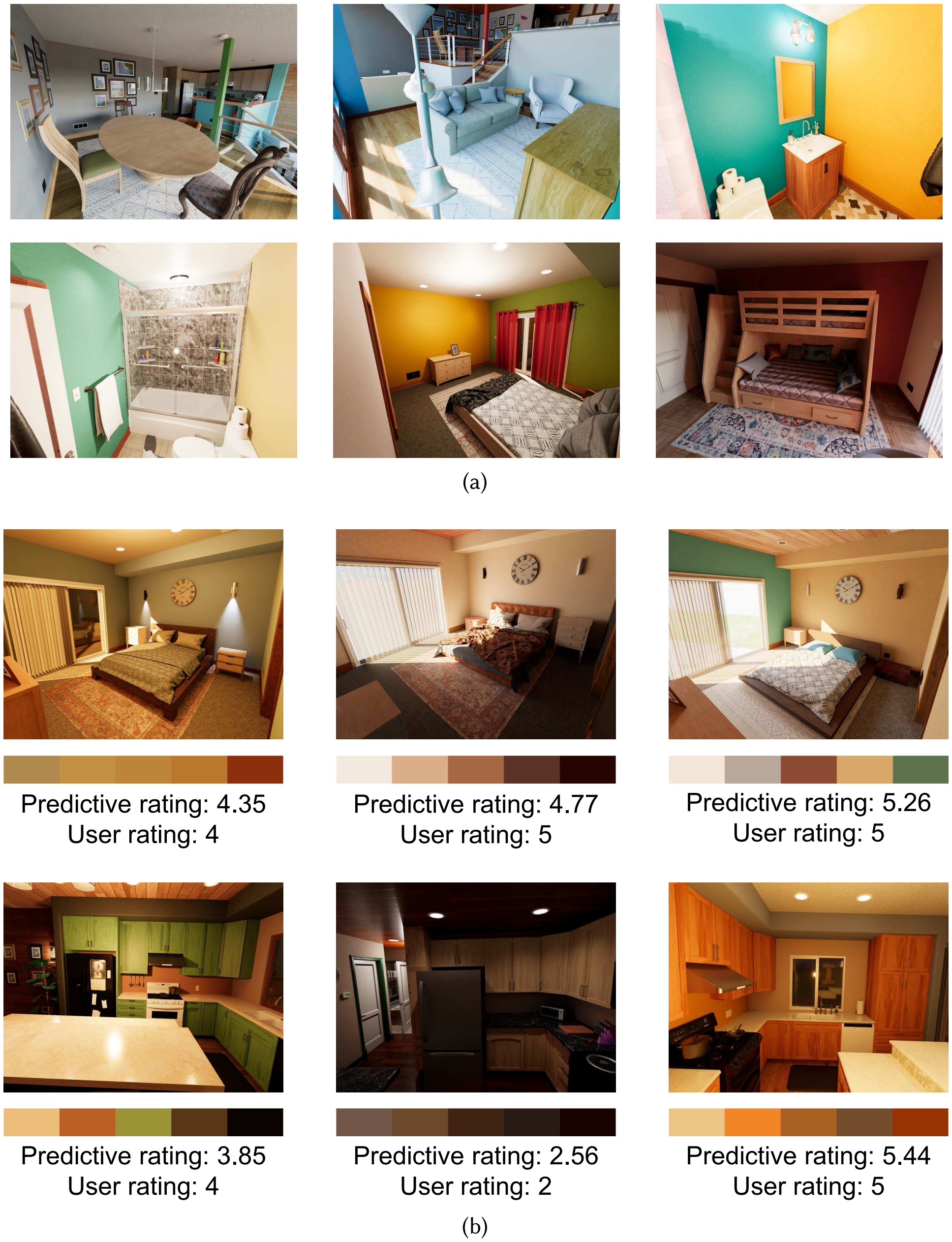


Fig. 21. (a) Preferred interior color designs for participant #13 in different living spaces. (b) Top: User color preferences for interior design in the same room and layouts. Bottom: User color preferences for interior design the same function room but different layouts. Taking participant #13 as an example, below each image are accompanied by a comparison between the personalized predicted rating from our method and the preference rating from participant #13.

Table 11. Comparison of Predicted Ratings of Different Color Theme Evaluation Models and Users' Personalized Ratings in Personalized Fashion Outfit Recommendation

Metric / Model	MAE	RMSE	ρ_P	ρ_{SP}
Our method	**0.5635**	**0.5813**	**0.7471**	**0.7233**
Color pair based [68]	0.8782	0.8536	0.3327	0.2689
Collaborative filtering [41]	0.7281	0.6534	0.5533	0.5836

Bold fonts indicate the best results.

Table 12. Comparison of Predicted Ratings of Different Color Theme Evaluation Models and Users' Personalized Ratings in Interior Design Color Recommendation

Metric / Model	MAE	RMSE	ρ_P	ρ_{SP}
Our method	**0.5063**	**0.5308**	**0.7735**	**0.7564**
Color pair based [68]	0.8080	0.8336	0.3658	0.3125
Collaborative filtering [41]	0.7781	0.6934	0.5233	0.5035

Bold fonts indicate the best results.

We continued to select 30 participants (15 males and 15 females, aged 20–30 years) without professional visual perception or color science experience to rate the color themes. Our method calculated the aesthetic preference of each participant and predicted aesthetic preference ratings for the test images. Table 12 presents a comparison of color ratings predicted by our method against those from the state-of-the-art methods, demonstrating that our quantitative assessment outperforms other methods. Figure 21(a) displays the interior color design favored by participant #13 for each living space. Our method can also select color schemes that match user preferences in similar interior scenes. The first row of Figure 21(b) shows interior scenes with different color pairings but the same room and layouts. Below each image are the predicted rating provided by our method and the preference rating given by participant #13. The results indicate that our method accurately simulates user color preferences and assists users in selecting interior scene designs that better match their color preferences. The second row of Figure 21(b) demonstrates interior scene designs for the same living space (kitchen) in different layouts. Despite the variations in room layouts, our method continues to accurately capture user color preferences, suggesting that our approach can assist users in choosing their preferred color decoration styles.

7 Discussion and Conclusion

We present a new color theme evaluation method based on the different aesthetic cognition of individuals. To the best of our knowledge, our work is the first-of-its-kind method to explicitly model and incorporate aesthetic cognition of individuals for color theme evaluation. Experimental results show that our method outperforms state-of-the-art methods for various test cases. Since our model relies on the aesthetic preferences of different users, if fewer than five users make comments or give ratings on color themes, we may predict less accurate results. Our current model is a preliminary exploration of color theme evaluation with user aesthetic awareness. We plan to further design advanced algorithms and systems to facilitate novices to cope with various real-world application scenarios.

References

[1] Adobe COLOR CC. 2019. Retrieved from https://color.adobe.com/

[2] COLORLOVERS. 2019. Retrieved from http://www.colourlovers.com/

[3] Colormind. 2019. Retrieved from http://colormind.io/

[4] A. Baszczynska. 2016. Kernel estimation of cumulative distribution function of a random variable with bounded support. *Statistics in Transition New* 17, 3 (2016), 541–556.

[5] Erik Cambria, Yang Li, Frank Z. Xing, Soujanya Poria, and Kenneth Kwok. 2020. SenticNet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 105–114.

[6] Ying Cao, Antoni B. Chan, and Rynson W. H. Lau. 2017. Mining probabilistic color palettes for summarizing color use in artwork collections. In *Proceedings of the SIGGRAPH Asia 2017 Symposium on Visualization*. 1–8.

[7] J. Douglas Carroll and Phipps Arabie. 1998. Multidimensional scaling. In *Measurement, Judgment and Decision Making*, Michael H. Birnbaum (Ed.), Academic Press, San Diego, 179–250.

[8] Christel Chamaret. 2016. *Color Harmony: Experimental and Computational Modeling*. Ph.D. Dissertation. Université Rennes 1.

[9] Huiwen Chang, Ohad Fried, Yiming Liu, Stephen DiVerdi, and Adam Finkelstein. 2015. Palette-based photo recoloring. *ACM Transactions on Graphics* 34, 4 (2015), 139–1.

[10] Xiaowu Chen, Dongqing Zou, Jianwei Li, Xiaochun Cao, Qinping Zhao, and Hao Zhang. 2014. Sparse dictionary learning for edit propagation of high-resolution images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2854–2861.

[11] Yiran Chen, Pengfei Liu, Ming Zhong, Zi-Yi Dou, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. CDEvalSumm: An empirical study of cross-dataset evaluation for neural summarization systems. DOI: https://doi.org/10.48550/arXiv.2010.05139

[12] Gilbert A. Churchill Jr and J. Paul Peter. 1984. Research design effects on the reliability of rating scales: A meta-analysis. *Journal of Marketing Research* 21, 4 (1984), 360–375.

[13] Daniel Cohen-Or, Olga Sorkine, Ran Gal, Tommer Leyvand, and Ying-Qing Xu. 2006. Color harmonization. In *Proceedings of the ACM Transactions on Graphics (TOG)*, Vol. 25. ACM, 624–630.

[14] John Dawes. 2008. Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. *International Journal of Market Research* 50, 1 (2008), 61–104.

[15] Khosrow Dehnad. 1987. Density estimation for statistics and data analysis. *Technometrics* 29, 4 (1987), 495–495.

[16] Connor C. Gramazio, David H. Laidlaw, and Karen B. Schloss. 2016. Colorgorical: Creating discriminable and preferable color palettes for information visualization. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (2016), 521–530.

[17] Xintong Han, Zuxuan Wu, Yu-Gang Jiang, and Larry S. Davis. 2017. Learning fashion compatibility with bidirectional LSTMs. In *Proceedings of the 25th ACM International Conference on Multimedia*. 1078–1086.

[18] Yu Han, Chen Xu, George Baciu, Min Li, and Md Robiul Islam. 2016. Cartoon and texture decomposition-based color transfer for fabric images. *IEEE Transactions on Multimedia* 19, 1 (2016), 80–92.

[19] Anya C. Hurlbert and Yazhu Ling. 2007. Biological components of sex differences in color preference. *Current biology* 17, 16 (2007), R623–R625.

[20] Johannes Itten. 1970. *The Elements of Color*. John Wiley & Sons.

[21] Ankur Joshi, Saket Kale, Satish Chandel, and D. Kumar Pal. 2015. Likert scale: Explored and explained. *Current Journal of Applied Science and Technology* 7, 4 (2015), 396–403.

[22] Naoki Kita and Kazunori Miyata. 2016. Aesthetic rating and color suggestion for color palettes. In *Proceedings of the Computer Graphics Forum*, Vol. 35. Wiley Online Library, 127–136.

[23] Guoliang Li, Jiannan Wang, Yudian Zheng, and Michael J. Franklin. 2016. Crowdsourced data management: A survey. *IEEE Transactions on Knowledge and Data Engineering* 28, 9 (2016), 2296–2319.

[24] Leida Li, Hancheng Zhu, Sicheng Zhao, Guiguang Ding, and Weisi Lin. 2020. Personality-assisted multi-task learning for generic and personalized image aesthetics assessment. *IEEE Transactions on Image Processing* 29 (2020), 3898–3910.

[25] Xujie Li, Hanli Zhao, Guizhi Nie, and Hui Huang. 2015. Image recoloring using geodesic distance based color harmonization. *Computational Visual Media* 1, 2 (2015), 143–155.

[26] Sharon Lin and Pat Hanrahan. 2013. Modeling how people extract color themes from images. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 3101–3110.

[27] Sharon Lin, Daniel Ritchie, Matthew Fisher, and Pat Hanrahan. 2013. Probabilistic color-by-numbers: Suggesting pattern colorizations using factor graphs. *ACM Transactions on Graphics (TOG)* 32, 4 (2013), 37.

[28] Yazhu Ling and Anya C. Hurlbert. 2007. A new model for color preference: Universality and individuality. In *Final Program and Proceedings-IS and T/SID Color Imaging Conference*. Newcastle University, 8–11.

[29] Shiguang Liu and Huarong Luo. 2016. Hierarchical emotional color theme extraction. *Color Research & Application* 41, 5 (2016), 513–522.

[30] Peng Lu, Zhijie Kuang, Xujun Peng, and Ruifan Li. 2014. Discovering harmony: A hierarchical colour harmony model for aesthetics assessment. In *Proceedings of the Asian Conference on Computer Vision*. Springer, 452–467.

[31] Peng Lu, Xujun Peng, Xinshan Zhu, and Ruifan Li. 2016. An EL-LDA based general color harmony model for photo aesthetics assessment. *Signal Processing* 120 (2016), 731–745.

[32] Pei Lv, Meng Wang, Yongbo Xu, Ze Peng, Junyi Sun, Shimei Su, Bing Zhou, and Mingliang Xu. 2018. USAR: An interactive user-specific aesthetic ranking framework for images. In *Proceedings of the 26th ACM international conference on Multimedia*. 1328–1336.

[33] Yutaka Matsuda. 1995. Color design. *Asakura Shoten* 2, 4 (1995), 10.

[34] John Maule, Alice E. Skelton, and Anna Franklin. 2023. The development of color perception and cognition. *Annual Review of Psychology* 74 (2023), 87–111.

[35] Saif M. Mohammad and Felipe Bravo-Marquez. 2017. Emotion intensities in tweets. DOI: https://doi.org/10.48550/arXiv.1708.03696

[36] Parry Moon and Domina Eberle Spencer. 1944. Geometric formulation of classical color harmony. *Journal of the Optical Society of America A* 34, 1 (1944), 46–59.

[37] Naila Murray, Luca Marchesotti, and Florent Perronnin. 2012. AVA: A large-scale database for aesthetic visual analysis. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2408–2415.

[38] Jerome L. Myers, Arnold Well, and Robert Frederick Lorch. 2010. *Research Design and Statistical Analysis*. Routledge.

[39] Masashi Nishiyama, Takahiro Okabe, Imari Sato, and Yoichi Sato. 2011. Aesthetic quality classification of photographs based on color harmony. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR '11)*. IEEE, 33–40.

[40] Peter O'Donovan, Aseem Agarwala, and Aaron Hertzmann. 2011. Color compatibility from large datasets. *ACM Transactions on Graphics* 30, 4, Article 63 (July 2011), 12 pages.

[41] Peter O'Donovan, Aseem Agarwala, and Aaron Hertzmann. 2014. Collaborative filtering of color aesthetics. In *Proceedings of the Workshop on Computational Aesthetics*. ACM, 33–40.

[42] Li-Chen Ou, Patrick Chong, M. Ronnier Luo, and Carl Minchew. 2011. Additivity of colour harmony. *Color Research & Application* 36, 5 (2011), 355–372.

[43] Li-Chen Ou and M. Ronnier Luo. 2006. A colour harmony model for two-colour combinations. *Color Research & Application* 31, 3 (2006), 191–204.

[44] Stephen E. Palmer and William S. Griscom. 2013. Accounting for taste: Individual differences in preference for harmony. *Psychonomic Bulletin & Review* 20 (2013), 453–461.

[45] Stephen E. Palmer and Karen B. Schloss. 2010. An ecological valence theory of human color preference. *Proceedings of the National Academy of Sciences* 107, 19 (2010), 8877–8882.

[46] Stephen E. Palmer, Karen B. Schloss, and Jonathan Sammartino. 2013. Visual aesthetics and human preference. *Annual Review of Psychology* 64 (2013), 77–107.

[47] Huy Q. Phan, Hongbo Fu, and Antoni B. Chan. 2017. Color orchestra: Ordering color palettes for interpolation and prediction. *IEEE Transactions on Visualization and Computer Graphics* 24, 6 (2017), 1942–1955.

[48] Jian Ren, Xiaohui Shen, Zhe Lin, Radomir Mech, and David J. Foran. 2017. Personalized image aesthetics. In *Proceedings of the IEEE International Conference on Computer Vision*. 638–647.

[49] Kirk Roberts, Michael A. Roach, Joseph Johnson, Josh Guthrie, and Sanda M. Harabagiu. 2012. EmpaTweet: Annotating and detecting emotions on Twitter. In *Proceedings of the Language Resources and Evaluation Conference*, Vol. 12. Citeseer, 3806–3813.

[50] Miho Saito. 1996. Comparative studies on color preference in Japan and other Asian regions, with special emphasis on the preference for white. *Color Research & Application* 21, 1 (1996), 35–49.

[51] Karen B. Schloss, Laurent Lessard, Chris Racey, and Anya C. Hurlbert. 2018a. Modeling color preference using color space metrics. *Vision Research* 151 (2018), 99–116.

[52] Karen B. Schloss, G. V. P. L. MacDonald, and C. P. Biggam. 2018b. A color inference framework. In *Progress in Colour Studies: Cognition, Language and Beyond*, 107–122.

[53] Karen B. Schloss, Rolf Nelson, Laura Parker, Isobel A. Heck, and Stephen E. Palmer. 2017. Seasonal variations in color preference. *Cognitive Science* 41, 6 (2017), 1589–1612.

[54] Karen B. Schloss and Stephen E. Palmer. 2011. Aesthetic response to color combinations: preference, harmony, and similarity. *Attention, Perception, & Psychophysics* 73, 2 (2011), 551–571.

[55] Karen B. Schloss, Rosa M. Poggesi, and Stephen E. Palmer. 2011. Effects of university affiliation and "school spirit" on color preferences: Berkeley versus Stanford. *Psychonomic Bulletin & Review* 18 (2011), 498–504.

[56] Maria Shugrina, Jingwan Lu, and Stephen Diverdi. 2017. Playful palette: An interactive parametric color mixer for artists. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 61.

[57] Maria Shugrina, Wenjia Zhang, Fanny Chevalier, Sanja Fidler, and Karan Singh. 2019. Color builder: A direct manipulation interface for versatile color theme authoring. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[58] KyoungHee Son, Seo Young Oh, Yongkwan Kim, Hayan Choi, Seok-Hyung Bae, and Ganguk Hwang. 2015. Color sommelier: Interactive color recommendation system based on community-generated color palettes. In *Adjunct Proceedings of ACM Symposium on User Interface Software & Technology*. 95–96.

[59] Eli D. Strauss, Karen B. Schloss, and Stephen E. Palmer. 2013. Color preferences change after experience with liked/disliked colored objects. *Psychonomic Bulletin & Review* 20 (2013), 935–943.

[60] Yosephine Susanto, Andrew G. Livingstone, Bee Chin Ng, and Erik Cambria. 2020. The hourglass model revisited. *IEEE Intelligent Systems* 35, 5 (2020), 96–102.

[61] Jianchao Tan, Jose Echevarria, and Yotam Gingold. 2018. Efficient palette-based decomposition and recoloring of images via RGBXY-space geometry. In *SIGGRAPH Asia 2018 Technical Papers*. ACM, 262.

[62] Zhen Tang, Zhenjiang Miao, Yanli Wan, and Zhifei Wang. 2011. Color harmonization for images. *Journal of Electronic Imaging* 20, 2 (2011), 023001.

[63] Chloe Taylor, Karen Schloss, Stephen E. Palmer, and Anna Franklin. 2013. Color preferences in infants and adults are different. *Psychonomic Bulletin & Review* 20 (2013), 916–922.

[64] Unity Technologies. 2022. Unity SynthHomes: A Synthetic Home Interior Dataset Generator. Retrieved from https://github.com/Unity-Technologies/SynthHomes

[65] Primož Weingerl and Dejana Javoršek. 2018. Theory of colour harmony and its application. *Tehnički vjesnik* 25, 4 (2018), 1243–1248.

[66] Stephen Westland, Kevin Laycock, Vien Cheung, Phil Henry, and Forough Mahyar. 2007. Colour harmony. *Colour: Design & Creativity* 1, 1 (2007), 1–15.

[67] Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier R. Movellan, and Paul L. Ruvolo. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Proceedings of the Advances in Neural Information Processing Systems*. 2035–2043.

[68] Bailin Yang, Tianxiang Wei, Xianyong Fang, Zhigang Deng, Frederick WB Li, Yun Ling, and Xun Wang. 2019. A color-pair based approach for accurate color harmony estimation. In *Proceedings of the Computer Graphics Forum*, Vol. 38. Wiley Online Library, 481–490.

[69] Kazuhiko Yokosawa, Karen B. Schloss, Michiko Asano, and Stephen E. Palmer. 2016. Ecological effects in cross-cultural differences between US and Japanese color preferences. *Cognitive Science* 40, 7 (2016), 1590–1616.

[70] Qing Zhang, Chunxia Xiao, Hanqiu Sun, and Feng Tang. 2017. Palette-based image recoloring using color decomposition optimization. *IEEE Transactions on Image Processing* 26, 4 (2017), 1952–1964.

[71] Hancheng Zhu, Leida Li, Jinjian Wu, Sicheng Zhao, Guiguang Ding, and Guangming Shi. 2020. Personalized image aesthetics assessment via meta-learning with bilevel gradient optimization. *IEEE Transactions on Cybernetics* 52, 3 (2020), 1798–1811.

[72] Xingxing Zou, Kaicheng Pang, Wen Zhang, and Waikeung Wong. 2022. How good is aesthetic ability of a fashion model? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21200–21209.

Received 27 July 2023; revised 15 April 2024; accepted 22 April 2024